UNIVERSITAT POLIÈCNICA DE CATALUNYA (UPC) - BARCELONATECH

FACULTAT D'INFORMÀTICA DE BARCELONA (FIB)

MASTER EN INGENIERÍA INFORMTÍCA

Interpretación automática de clases a partir de la extensión del cuadro termómetro a variables cualitativas a través de KLASS

Trabajo de fin de máster

Autor:
Johnny Javier Ávila Montalvo.

Directora:
Karina Gibert Oliveras.

Departamento:
Estadística i Investigació Operativa.

Fecha de presentación:
25 de junio de 2018.

Resumen

En este proyecto se describe el proceso de KDD con un énfasis principal en la fase de interpretación de los resultados y generación del conocimiento. Para esto, en primer lugar hace una revisión del estado del arte y se presentan tres herramientas que ayudan a la interpretación de las clases descubiertas al ejecutar un proceso de minería de datos como son el CPG, el TLP y el aTLP. Y luego el termómetro como una herramienta que permite introducir el conocimiento a priori de los expertos para transferir la polaridad semántica de las variables al TLP. Luego se explica brevemente del proceso de KDD y se introduce la herramienta Java-Klass como un software que brinda un conjunto de herramientas para ayudar a los expertos en el proceso de minería de datos.

A continuación, este documento se enfoca principalmente en describir el modelo de termómetro propuesto originalmente en [Canudes Solans, 2016] para variables numéricas y el proceso de generación automática del TLP a partir de este; siguiendo con esta línea, se propone un modelo visual y estructural para extender los termómetros a variables cualitativas así como las modificaciones al proceso de generación automática del TLP para que sea posible generarlo con el termómetro extendido y las modificaciones realizadas en java para que el sistema trabaje con los termómetros para los dos tipos de variables y sin perder la funcionalidad original. Al final se hace un análisis de un caso real de aplicación de esta herramienta sobre datos relativos a un proyecto desarrollado para la OMS sobre los sistemas de salud mental en países en vías de desarrollo, así como una comparación entre el TLP generado a partir del termómetro extendido con el que en su día propusieron los expertos y se concluye que el termómetro generado aproxima bastante bien al original.

Palabras Clave: termómetros, semáforos, interpretación, KDD, minería de datos.

Johnny Javier Ávila Montalvo.

Agradecimientos

Es justo expresar mis más sinceros agradecimientos a Karina Gibert, Ph.D por el tiempo y esfuerzo dedicados a la tutoría de esta tesis de máster, por su apoyo absoluto e incondicional, su inagotable paciencia y por sus consejos, enseñanzas y correcciones brindados desde un principio y a lo largo de toda la realización de este proyecto así como en la asignatura de "Técnicas de Minería de Datos". Agradezco también a Carlos Jordán, Luis Pérez, Lavanya Mandadapu y Beatriz Sevilla por el apoyo brindado como compañeros de desarrollo de Java-KLASS. Asimismo, reconozco y agradezco a la Secretaría de Educación Superior, Ciencia, Tecnología e Innovación del gobierno del Ecuador por ser la entidad auspiciaste de la beca para la realización de mis estudios de Máster y el presente proyecto.

Índice general

Resumen II
Índice general IV
Índice de figuras VI
Índice de tablas VIII

1. Introducción **1**
1.1. Contexto 1
1.2. Motivación 2
1.3. Objetivos 3
1.3.1. Objetivo General 3
1.4. Trabajos Previos y estado del arte 4
1.4.1. Trafic Ligth Panels 4
1.4.2. Annotated Trafic Ligth Panels 5
1.4.3. Termómetros 7
1.5. Organización del documento 7

2. Metodología y Antecedentes **9**
2.1. El proceso de KDD 9
2.1.1. Comprender el dominio y definir las metas de la aplicación . 10
2.1.2. Creación del dataset objetivo 10
2.1.3. Limpieza y Pre-proceso de los datos 11
2.1.4. Transformación de los datos 13
2.1.5. Minería de datos (Data minig) 14
2.1.6. Interpretación de los resultados 15
2.1.7. Uso del conocimiento descubierto 15
2.2. Introducción a Java-KLASS 15
2.2.1. Funcionalidades de Java-KLASS 16
2.2.2. Cronología 17
2.3. Trabajo Realizado 23

2.3.1. Introducción 23
2.3.2. Antecedentes 23
2.3.3. Desarrollo del termómetro para variables cualitativas 30
2.3.4. Generación del TLP a partir de los termómetros para variables cualitativas 35
2.3.5. Fomalización del proceso generalizado 39

3. Caso de Estudio **42**
3.1. Introducción 42
3.2. Descripción de los datos de la OMS 42
3.3. Clasificación o *profiling* de los sistemas de salud mental 43
3.4. Aplicación del termómetro para variables numéricas 51
3.5. Aplicación del termómetro para variables cualitativas 53
3.6. Creación de los termómetros en Java-KLASS 54
3.7. Construcción del TLP a partir del termómetro completo 56
3.8. Comparación con el TLP original 60
3.8.1. Comparación entre TLPs anotados 63
3.8.2. Verificación de los perfiles a partir del TLP generado 64
3.9. Discusión 65

4. Conclusiones **67**
4.1. Conclusiones 67
4.2. Futuras líneas de investigación 69

Acrónimos **71**

Bibliografía. **72**

Índice de figuras

1.1. Ejemplo de un Class Panel Graph 5
1.2. Modelo de gradación de los colores. Fuente [Gibert and Conti, 2015] 6

2.1. Esquema completo del proceso de KDD. Adaptado de [Fayyad et al., 1996] 10
2.2. Ejemplo de un outlier en la combinación de 2 variables. **Fuente:** *https://madhureshkumar.files.wordpress.com/2015/06/multivariate-outlier-example.jpg* 13
2.3. Cronología de Klass. Parte 1 21
2.4. Cronología de Klass. Parte 2 22
2.5. Diseño del termómetro. Fuente: [Canudes Solans, 2016] 24
2.6. Termómetro en Java-KLASS. 25
2.7. Termómetro en Java-KLASS. 27
2.8. Asociación de colores. 29
2.9. Diseño del termómetro para variables cualitativas. 31
2.10. Termómetro con variables cualitativas en Java-KLASS 31
2.11. Termómetro con variables combinadas en Java-KLASS 32
2.12. Archivo CSV con termómetros para variables cuantitativas y cualitativas 33
2.13. Diseño original del diagrama de clases de los termómetros 33
2.14. Diseño original del diagrama de clases de los termómetros 34

3.1. CAJ: Árbol general de clasificación tallado en 7 clases. 44
3.2. Mapa de los países con su clase correspondiente. 45
3.4. TLP creado manualmente con ayuda de los expertos 46
3.3. CPG con las clases descubiertas y las variables para la interpretación. 47
3.5. Modelo conceptual de los termómetros para variables numéricas. **Fuente:** [Canudes Solans, 2016] 52
3.6. Termómetro creado en Java-KLASS. 52

3.7. TLP generado a partir de los termómetros para variables numéricas. 53
3.8. Modelo conceptual de los termómetros para variables cualitativas. . 54
3.9. Termómetros creados en Java-KLASS 55
3.10. Fase de recodificación/discretización 56
3.11. Tablas cruzadas 57
3.12. TLP generado 58
3.13. TLP generado a partir de los termómetros en Java-KLASS con $\gamma =$ 0,75 59
3.14. Análisis descriptivo por clases 60
3.15. Comparación de TLPs generados con diferentes configuraciones . . 62
3.16. Comparación del aTLP original con el generado con $\gamma = 0{,}85$ 64

Índice de tablas

2.1. Ejemplo de dataset. 11
2.2. Estructura de los termómetros numéricos. **Fuente**:[Canudes Solans, 2016] 26
2.3. Resultado de plicar la discretización sobre el termómetro de la figura 2.7 28
2.4. Ejemplo de tabla cruzada 28
2.5. Ejemplo de TLP 30
2.6. Estructura de los termómetros para variables cualitativas 32
2.7. Aplicación del algoritmo de asignación de color con un valor de $\gamma = 0{,}6$ 38

3.1. Clasificación de los servicios de salud mental en los distintos países . 45
3.2. Descripción de las variables usadas para la interpretación 46
3.3. Tabla de comparación de TLPs respecto al original basado en expertos 63

Capítulo 1

Introducción

1.1. Contexto

Los avances tecnológicos en la adquisición, almacenamiento y procesamiento de la información han resultado en un enorme crecimiento de las bases de datos en prácticamente todos los ámbitos, desde bases de datos de negocios, estadísticas gubernamentales, médicas, entre otros[Hand, 2007][Goebel and Gruenwald, 1999]; sin embargo, éste crecimiento ha hecho que hoy en día sea casi imposible su análisis manual para obtener conocimientos válidos que soporten a la toma de decisiones. En este ámbito, las técnicas de minería de datos brindan una herramienta poderosa para el descubrimiento del conocimiento a partir de los datos, dichas técnicas por lo general siguen un proceso completo de análisis previo de los datos, limpieza, imputación de errores, datos faltantes y corrección de formatos para luego proceder al descubrimiento de información utilizando métodos estadísticos, redes neuronales o métodos de aprendizaje automático. [Zhu, 2007] Finalmente se realiza un post proceso de la información encontrada para obtener el conocimiento útil para la toma de decisiones.

A pesar de que en el proceso de minería de datos y descubrimiento de la información se pueden automatizar algunas etapas, en la mayoría de las áreas o aplicaciones reales prácticas, es muy importare la intervención humana para enriquecer los datos con conocimiento externo brindado por los expertos en cada área de estudio[Prather et al., 1997]. Básicamente el experto ejerce un papel fundamental en las siguientes etapas: agregar información útil o bases de conocimiento a priori, para dar contexto a los datos y para interpretar los resultados obtenidos al aplicar una técnica específica[Zhu, 2007].

En este contexto es importante contar con herramientas que ayuden a incorporar conocimiento a priori del experto al sistema, así como a interpretar los resultados obtenidos. Este proyecto propone una intervención sobre el software Klass que contribuye a facilitar esta tarea poco estudiada. En el capítulo 2.2 se describe Klass como un software de soporte a los procesos integrales de data science e incorpora un módulo específico de interpretación automática de clases que incluye entre otros la construcción de semáforos a partir de las distribuciones condicionadas de las variables. Por otro lado, el termómetro es una herramienta de transferencia de la semántica de las variables al sistema que funciona actualmente para variables numéricas. Tanto el semáforo como el termómetro se alimentan con información brindada por los expertos en el área de estudios. El proyecto extenderá el modelo del termómetro a las variables cualitativas y cerrará el ciclo de interpretación de clases trasladando la información semántica del termómetro a la construcción de semáforos bajo distintos esquemas de conexión entre las dos herramientas.

1.2. Motivación

En [Gibert et al., 2008a, Gibert and Conti, 2015] se describen los semáforos o en inglés *Traffic Ligth Panel* (TLP) como una herramienta simbólica de mucha utilidad para el post-proceso de los resultados de un análisis cluster. Los TLP asocian los colores de un semáforo con las tendencias de las variables de cada clase para ayudar al entendimiento y la conceptualización de las clases descubiertas. Asimismo, [Canudes Solans, 2016] presenta el termómetro como una herramienta para modelar el conocimiento previo de los expertos permitiendo inyectar al sistema información relativa a la interpretación de las variables desde un punto de vista semántico y manteniendo el modelo visual de los TLP. A partir de un termómetro se puede clasificar el rango de los valores de una variable cuantitativa en intervalos de tres zonas de colores (verde, amarillo y rojo) donde el rojo representa valores problemáticos o poco adecuados de la variable en el problema estudiado de acuerdo al conocimiento de los expertos, y el verde se asociará a los valores más benévolos. De igual forma, aquí se presenta un método para transferir el conocimiento a priori introducido en los termómetros a las herramientas de conceptualización de las clases. De esta forma se automatiza la interpretación del cluster en términos de la semántica original de las variables.

Los termómetros presentados en [Canudes Solans, 2016] sólo permiten transferir la semántica de variables numéricas, sin embargo, como se muestra en [Gibert and Cortés, 1997], es posible realizar un cluster con variables cuantitativas y cualitativas usando técnicas mixtas, por lo que los TLP ya permiten trabajar con estos dos tipos de variable, por lo tanto, es importante extender los termómetros para que sea posible recodificar variables cualitativas y luego transferir este conocimiento a los TLP automatizando su construcción y completando el ciclo de interpretación de los resultados.

1.3. Objetivos

1.3.1. Objetivo General

En este trabajo se cubre un doble objetivo:

- Generalizar el módulo de termómetro presentado en [Canudes Solans, 2016] a variables cualitativas.
- Estudiar el impacto de utilizar la información de los termómetros en la construcción de los semáforos respecto a su formulación original basada en las distribuciones condicionadas de las variables respecto a las clases.

Para ello se afrontan los siguientes objetivos específicos:

1. Extender los termómetros a variables cuantitativas.
2. Crear un método de recodificación de variables cualitativas basada en termómetros que asigne un código a las modalidades de la variable según los colores del termómetro.
3. Flexibilizar la implementación del TLP basado en termómetros actual para que: a) admita termómetros que incluyan variables cualitativas, b) permita trabajar con termómetros que no incluyan todas las variables del TLP o puedan contener variables adicionales.
4. Comparar el TLP generado automáticamente a partir de los termómetros generalizados con el TLP que se crearía de forma manual con ayuda del experto a partir de las distribuciones condicionadas de las variables respecto a las clases.

1.4. Trabajos Previos y estado del arte

En esta sección se presenta un resumen de los trabajos realizados previamente y que han servido como base para la realización de este proyecto.

1.4.1. Trafic Ligth Panels

En [Gibert et al., 2008a] se presentan los TLP como una herramienta de gran utilidad para facilitar la interpretación de las clases resultantes al aplicar un proceso de clusterización. En este trabajo se usa esta herramienta en el estudio de un caso real para descubrir patrones de respuesta a los tratamientos de rehabilitación para pacientes con daño cerebral. Luego de aplicar un proceso de *Knowledge Discovery in Databases* (KDD)(Descubrimiento de conocimiento en bases de datos), y de introducir una base de conocimiento *Knowledge Base* (KB) a priori, el sistema recomienda 4 clases que luego de la interpretación, y desde el punto de vista médico, se corresponden con diferentes patrones en el nivel de incremento de respuesta al tratamiento. Con los resultados obtenidos se crea un *Class Panel Graph* (CPG) que consiste en una representación gráfica en forma de panel que contiene las variables en las columnas y las clases en las filas en el que se muestran las distribuciones condicionadas de las variables para cada clase en forma de histograma o de boxplots múltiples (como se muestra en el ejemplo de la figura 1.1). Al final se abstrae el CPG asociando colores a las casillas del CPG según su tendencia central y el significado de las variables y se transforma en un TLP para ayudar al experto en la conceptualización final de las clases. Con se evidencia el TLP como una herramienta muy útil para presentar los resultados de la clusterización a un experto sin conocimientos técnicos y facilitar la interpretación de las clases y su conceptualiziín fundamental para después asignar acciones o decisiones a cada clase.

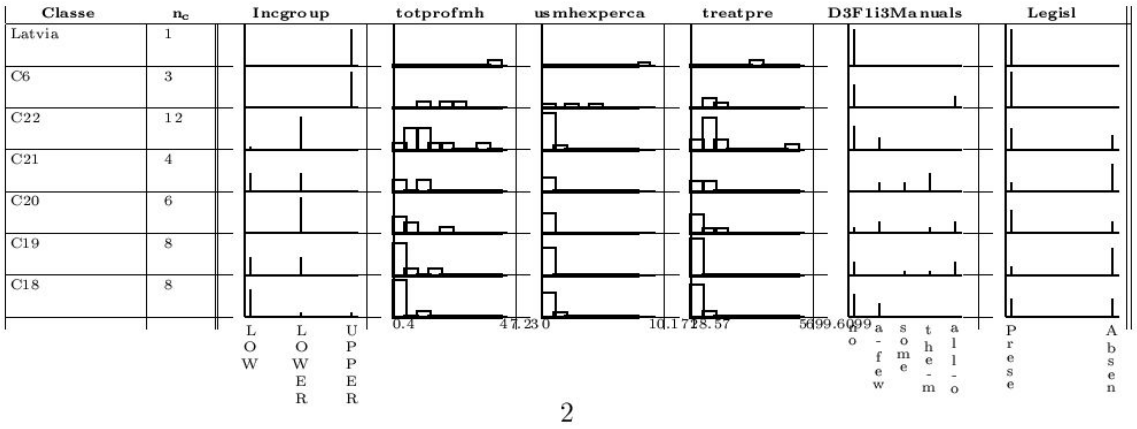

Figura 1.1: *Ejemplo de un Class Panel Graph*

En [Gibert et al., 2012b][Gibert et al., 2012a] se proponen métodos para automatizar la creación del TLP en los se divide el rango de cada variable en 3 intervalos que representan 3 niveles cualitativos y se corresponden con los códigos de color de los semáforos. La división de los intervalos es creada a partir de una tabla de semántica de polaridad explicada en el mismo artículo.

1.4.2. Annotated Trafic Ligth Panels

En [Gibert and Conti, 2015] se presenta una mejora para los TLP presentados anteriormente y se denomina *annotated Traffic Ligth Panel* (aTLP), los cuales sirven para gestionar la incertidumbre intrínseca que se produce cuando se interpretan los prototipos resultantes de una clusterización. El CPG se limitaba a representar la tendencia central de las variables en las clases pero el análisis de las variables asociadas se tenía que hacer con unas tablas de estadística básicas condicionadas a las clases con el propio software proporcionado que presenta entre otros las desviaciones típicas de cada variable en cada clase. Los aTLP asocian a los colores del semáforo con dos dimensiones del (tono y saturación) que sirven para medir la tendencia central y la pureza de los prototipos basándose en los *Coeficientes de Variación* (CV) y un modelo de incertidumbre. En este sentido los colores puros representan baja o nula variabilidad en la tendencia central de las variables, mientras que los colores más obscurecidos representan un incremento en la heterogeneidad y por consecuencia una perdida de fiabilidad en la toma de decisiones basadas en las casillas más oscuras del TLP. En [Gibert and Conti, 2015] se presenta un modelo de cálculo automático de la saturación de cada color calculado en base a los coeficientes de variación o al factor de incertidumbre de cada variable en cada clase. En

la figura 1.2 se muestra el modelo de gradación o pérdida de pureza de los colores para coeficientes de variación en un intervalo entre [0,1].

RED Scale

x	R	G	B	Color
0,00	255	0	0	
0,05	244	0	0	
0,10	233	0	0	
0,15	222	0	0	
0,20	212	0	0	
0,25	202	0	0	
0,30	192	0	0	
0,35	182	0	0	
0,40	173	0	0	
0,45	164	0	0	
0,50	155	0	0	
0,55	146	0	0	
0,60	138	0	0	
0,65	130	0	0	
0,70	122	0	0	
0,75	114	0	0	
0,80	107	0	0	
0,85	100	0	0	
0,90	93	0	0	
0,95	86	0	0	
1,00	80	0	0	

GREEN Scale

x	R	G	B	Color
0,00	0	255	0	
0,05	0	244	0	
0,10	0	233	0	
0,15	0	222	0	
0,20	0	212	0	
0,25	0	202	0	
0,30	0	192	0	
0,35	0	182	0	
0,40	0	173	0	
0,45	0	164	0	
0,50	0	155	0	
0,55	0	146	0	
0,60	0	138	0	
0,65	0	130	0	
0,70	0	122	0	
0,75	0	114	0	
0,80	0	107	0	
0,85	0	100	0	
0,90	0	93	0	
0,95	0	86	0	
1,00	0	80	0	

Proposed YELLOW Scale

x	R	G	B	Color
0,00	255	255	0	
0,05	255	244	0	
0,10	254	233	0	
0,15	253	222	0	
0,20	252	212	0	
0,25	251	202	0	
0,30	249	192	0	
0,35	247	182	0	
0,40	245	173	0	
0,45	242	164	0	
0,50	239	155	0	
0,55	236	146	0	
0,60	232	138	0	
0,65	227	130	0	
0,70	222	122	0	
0,75	217	114	0	
0,80	211	107	0	
0,85	204	100	0	
0,90	197	93	0	
0,95	189	86	0	
1,00	180	80	0	

Figura 1.2: *Modelo de gradación de los colores. Fuente [Gibert and Conti, 2015]*

Para esto se utiliza el modelo RGB que representa los colores como un vector de 3 componentes, descomponiendo los colores en cantidades de rojo, verde y amarillo, cada componente toma valores entre 0 y 255 en donde 0 representa la ausencia del color y 255 la presencia del color con la máxima saturación. Como ya se ha dicho, en el aTLP los colores puros representan CV nulos, mientras que para los colores más obscuros el $CV = 1$. Este sería el valor de entrada de la columna x de la figura 1.2 y permitiría determinar el vector RBG de una clase Para calcular el grado de de-saturación de un color de acuerdo al coeficiente de variación se usan las siguieres ecuaciones:

- Para los colores rojo y verde(Primarios): $S(x) = 80 + 125(1-x) + 50(1-x)^2$
- Para el amarillo (Color compuesto): $S'(x) = 180 + 180(1-x) - 143(1-x)^2 + 38(1-x)^3$

Por lo tanto, para cada celda en el TLP en donde las filas representan las clases C de la variable categórica de clase, y las columnas la variable asociada X_k, el color de la celda es denotado como h_{Ck} y es expresado con el modelo RGB de la siguiente forma:

- $h_{Ck}(x) = (S(CV_{X_k}|C), 0, 0)$ para las celdas con valores no favorables o rojos.
- $h_{Ck}(x) = (0, S(CV_{X_k}|C), 0)$ para las celdas con valores favorables o verdes.

- $h_{Ck}(x) = (S'(CV_{X_k}, S(CV_{X_k}|C), 0)$ para las celdas con valores neutrales o amarillo.

Para determinar el valor de x se usa un identificador de la variación interna de la clase como se indica en [Gibert and Conti, 2015]. Si X_k es numérica $x = CV|C$ siendo C la clase de la que se quiere calcular el coeficiente de variación. Si X_k es cualitativa, se corresponde a la propagación de valores distintos a la frecuencia dominante de la clase. Es decir, para una clase identificada mayoritariamente como mujeres, se correspondería a la proporción de hombres que contiene la clase.

1.4.3. Termómetros

En [Canudes Solans, 2016] introduce el uso del termómetro como una herramienta de adquisición de conocimiento experto relacionado con la interpretación de la polaridad semántica de cada variable. En la sección 2.3.2 se presenta el formalismo del termómetro con detalle. En [Canudes Solans, 2016], se propone un modelo de termómetros en el que el rango de las variables cuantitativas es dividido en los 3 intervalos descritos de forma que el primer intervalo se corresponde con los valores más próximos a los mínimos de la variable, el segundo intervalo para los valores intermedios y el tercer intervalo con los valores más próximos a los máximos de la variable. El usuario transmite al sistema a través de esta herramienta visual en qué sentido han de interpretarse los valores extremos de las variables numéricas. En línea con los trabajos de automatización del TLP en [Gibert et al., 2012b][Gibert et al., 2012a], en [Canudes Solans, 2016] se propone un modelo para transferir la semántica de las variables expresada en un termómetro al semáforo con la posibilidad de asignar el color verde o rojo del semáforo tanto al primero como al último intervalo, de acuerdo al criterio del experto. Al final, se describe un método para construir automáticamente el TLP a partir de la información almacenada en el termómetro, relativo exclusivamente a variables numéricas, y se expone un caso práctico de construcción automática del TLP en el que se concluye que el resultado obtenido es considerado correcto por los expertos y que Klass reproduce el semáforo de forma fidedigna.

1.5. Organización del documento

El documento actual estará compuesto por la siguiente estructura:

- **Capitulo 1:** ***Introducción.*** Contexto, motivación, estado del arte y objetivos.
- **Capitulo 2:** ***Metodología y Antecedentes*** El proceso de KDD, introducción a Java-KLASS, antecedentes y trabajo realizado.
- **Capitulo 3:** ***Caso de estudio*** Ejemplo de una aplicación práctica en donde se clasifican datos de la *Organización Mundial de la Salud* (OMS) relativos a los sistemas de salud mental.
- **Capitulo 4:** ***Conclusiones y Resultados Finales.*** Evaluación de resultados, argumentos de conclusión y trabajos futuros.
- **Acrónimos** Una sección que incluye un listado de los acrónimos utilizados en la redacción del documento.
- ***Bibliografía.*** Se incluye toda la documentación utilizada.

Capítulo 2

Metodología y Antecedentes

En este capítulo se describe en primer lugar el marco teórico del proceso de KDD, luego se hace una breve introducción a la herramienta java-Klass y al final se describe el trabajo realizado en este proyecto.

2.1. El proceso de KDD

En [Leiva and S, 2018], se refieren al descubrimiento de conocimiento a partir de las bases de datos (KDD) como el proceso de *"identificar patrones patrones válidos, novedosos, potencialmente útiles y principalmente entendibles"*. En algunas ocasiones se usa el término minería de datos o *data mining* para referirse a esta actividad, sin embargo, la minería de datos es tan solo una etapa en la cual se aplican algoritmos para descubrir modelos o patrones dentro del proceso de KDD[Fayyad et al., 1996, Gibert et al., 2012b]. Recientemente a KDD se ha insertado en un marco más referencial que incorpora el apoyo a la toma de decisiones como parte del proceso y que se ha dado a llamar *Ciencia de Datos*[Gibert et al., 2018]. El proceso de KDD comporta una serie de pasos que incluyen: un análisis previo de los datos, limpieza, selección de las variables apropiadas, transformación, introducción de conocimientos a priori, optimizaciones necesarias y una interpretación correcta de los resultados que redunda en la producción de conocimiento accionable. El proceso completo se puede ver en la figura 2.1.

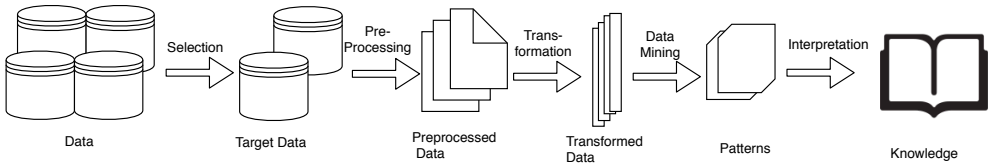

Figura 2.1: *Esquema completo del proceso de KDD. Adaptado de [Fayyad et al., 1996]*

Todos estos pasos se pueden ejecutar de forma iterativa con el fin de depurar o mejorar el proceso en sí. Cuando se habla de encontrar patrones válidos, se refiere a que el modelo debe funcionar para nuevos datos con cierta certeza[Fayyad et al., 1996]. De igual forma se requiere que los patrones sean novedosos, útiles y atendibles para el usuario final o el propietario[Leiva and S, 2018].

A continuación se dividirá el proceso de (KDD) en 7 pasos principales y se explicará sus funcionamiento y componentes.

2.1.1. Comprender el dominio y definir las metas de la aplicación

El primer paso en todo proceso de (KDD) aprender y entender el dominio de la aplicación y definir el alcance y las limitaciones del estudio a realizar. Luego, se debe recopilar toda la información y bases de conocimiento previas que puedan ser importantes. Como ya se ha mencionado anteriormente, el conocimiento previo es fundamental para enriquecer los datos. Aquí se debe reconocer las principales fuentes de información y quien las controla, de igual forma, se deben incluir los metadatos relacionados y dimensionar la cantidad de datos y formatos.

2.1.2. Creación del dataset objetivo

Un dataset es un conjunto de mediciones tomadas de algún ambiente o proceso. En otras palabras, se puede decir que un dataset es una colección de n objetos relacionados en donde cada objeto tiene un número de las mismas p mediciones, por lo tanto se puede construir una matriz de datos de $n*p$ en donde n representa el número de filas de la matriz o de mediciones tomadas en un ámbito (pacientes médicos, clientes de un banco, días de mediciones de un fenómeno, etc). Cada fila puede ser referida como un individuo, un caso, una entidad, un objeto, dependiendo del caso de estudio. Por otra parte, la p representa el número de mediciones hechas

sobre cada objeto o las columnas de la matriz, dependiendo del contexto de estudio se las suele llamar, variables, atributos o campos[Hand, 2007]. En la tabla 2.1 se muestra un ejemplo de básico de dataset.

ID	Edad	Educación	Sexo	Ocupación	Ingresos
145	45	Grado Universidad	M	Médico	45000
23	??	Instituto	M	Camarero	10000
718	20	Secundaria	F	Estudiante	??
52	45	Máster	F	Ejecutivo	38000
415	45	Grado Universidad	M	Artista	30000
23	??	Instituto	M	Artesano	15000
612	45	Grado Universidad	F	Programador	35000
931	45	Doctorado	F	Docente	41500
441	45	??	M	Granero	18000
7	??	Máster	M	Ingeniero	41000

Tabla 2.1: *Ejemplo de dataset.*

Aunque por lo general el conjunto de mediciones sobre un objeto suele ser el mismo, esto no necesariamente ocurre siempre, por ejemplo a determinados pacientes de un hospital se les puede aplicar un diferente número y tipo de tests o diferentes tratamientos. Por otro lado, los datos requeridos para construir un dataset no siempre se encuentran en una única fuente, estos pueden estar dispersos entre bases de datos relacionales, documentos, correos electrónicos, bases de datos de transacciones, registros de procesos o fuentes menos típicas como clips de video, fotografías, bases de datos geográficas entre otros. Por lo que en este paso lo más importante es seleccionar e integrar los datos relevantes para el estudio de todas fuentes heterogéneas, y luego homogeneizar los formatos para facilitar su proceso y análisis[Leiva and S, 2018].

2.1.3. Limpieza y Pre-proceso de los datos

El dataset objetivo creado en el paso anterior, por lo general está incompleto y/o "sucio", es decir que contiene datos o valores de atributos faltantes (missings)y tienen ruido, inconsistencias, errores y datos aislados (outliers). Los datos "sucios" pueden alterar los modelos descubiertos produciendo resultados inválidos o no confiables. El objetivo de este paso es mejorar la calidad de los datos por es importante

usar el conocimiento previo para elegir estrategias correctas para eliminar o tratar los valores de atributos faltantes, los outliers y las inconsistencias.

Por lo que se refiere al preprocesamiento, los nuevos contextos de big data y sistemas complejos que se trata actualmente requieren de la combinación de un número importante de operaciones previas al análisis que se deben organizar correctamente para mantener la consistencia de la base de datos y asegurar el correcto análisis de los missings. En [Gibert et al., 2016] se aporta una propuesta metodológica para abordar el preprocesameinto de propósito general.

Valores faltantes o missings

Los missings están presentes en casi todos los datasets en las aplicaciones reales, esto puede pasar por diferentes razones y pueden tener diferente naturaleza. De igual forma se pueden encontrar missings de manera totalmente aleatoria, es decir que no siguen ningún patrón en los datos, estos pueden ser producidos por errores humanos, fallos temporales en sensores, valores faltantes forzados, entre otros; o pueden existir missings no aleatorios, que son producidos por causas identificables y se pueden producir ya sea porque los datos han sido eliminados a propósito, por datos que no han sido proporcionados, porque para ciertos valores no es posible obtener su medición, entre otros[Gibert et al., 2016].

El tratamiento de los missings en la actualidad puede ser automatizado por muchos software de minería de datos, sin embargo, es importante conocer que tipo de tratamiento van a recibir los datos faltantes y evaluar si es el correcto para los datos y para los objetivos del estudio. En muchos casos, principalmente cuando los missings no aparecen de forma aleatoria, una mala decisión tomada por el software de forma transparente puede tener graves consecuencias para el análisis.

Una mejor forma de tratar los missings es la imputación que consiste en una serie de técnicas de estimación para convertir los datos faltantes en información válida, las diferentes técnicas dependen de la naturaleza de la variable y están ampliamente documentadas en [Gibert et al., 2016].

En el caso que nos ocupa, se ha utilizado la técnica MIMMI [Gibert, 2014].

Datos aislados o outliers

"Los outliers son instancias con valores muy extremos en una o más variables" [Gibert et al., 2016, Gibert et al., 2008b], además, los outliers pueden tener cierta

dimensionalidad, es decir, puede que el valor de una variable en particular no represente un outlier pero al combinar 2 o más variables, la relación entre estas puede ser inusual, en la figura 2.2 se puede ver un ejemplo en donde se observa mejor lo explicado.

Dependiendo de la técnica de minería de datos aplicada al estudio, algunas pueden ser más o menos robustas a la precencia de outliers, para las menos robustas es fundamental su identificación y tratar de forma adecuada cada outlier, caso contrario el modelo descubierto puede estar alterado. El tratamiento para los outliers dependerá de su naturaleza y las técnicas están documentadas en [Gibert et al., 2016].

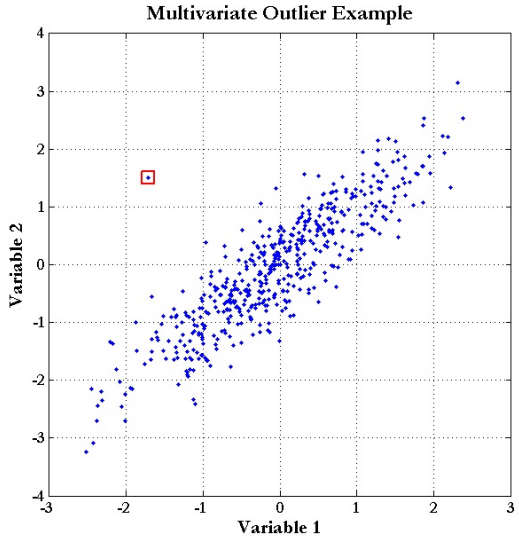

Figura 2.2: *Ejemplo de un outlier en la combinación de 2 variables.* ***Fuente:*** https://madhureshkumar.files.wordpress.com/2015/06/multivariate-outlier-example.jpg

2.1.4. Transformación de los datos

En algunos casos una o varias variables no están en el formato mas conveniente para realizar la minería de datos, por ejemplo al crear un dataset a partir de más de una fuente de datos, es posible que ciertos aspectos como unidades de medida o formas de representar una variable sean diferentes en ciertos grupos de observaciones, en estos casos es crucial homogeneizar las instancias. De igual forma se pueden hacer transformaciones para mejorar la interpretabilidad de los datos o porque el método de data mining así lo requiere, sin embargo, en [Gibert et al., 2016] sugieren que en ocasiones estas últimas transformaciones no son recomendadas, en lugar

de esto es mejor buscar una técnica de data mining que se acople mejor a la naturaleza de las variables. En [Gibert et al., 2016] se puede encontrar un estudio más profundo de las transformaciones que se pueden realizar a las variables o instancias del dataset.

2.1.5. Minería de datos (Data minig)

Luego de realizar las transformaciones necesarias, el siguiente paso es la minería de datos, en esta etapa se estudian con profundidad los datos para descubrir los patrones o modelos existentes, la herramienta fundamental para el data mining son los algoritmos. Esta etapa puede ser dividida a la ves en 3 pasos como se explica a continuación:

1. En primer lugar se debe definir el propósito del modelo requerido, es decir, se debe decidir si sobre los datos se requiere aplicar una clasificación, una regresión o algún otro método estadístico, o una clusterización.

2. El segundo paso consta en elegir el algoritmo apropiado que va a ser usado para el descubrimiento de patrones[Fayyad et al., 1996], en este paso se debe responder a las siguientes preguntas[Leiva and S, 2018]:

   - ¿Qué algoritmo es el más apropiados para buscar patrones a los datos?. Por ejemplo si se desea hacer una clasificación se puede elegir entre KNN o random forest.
   - ¿Cuáles son los parámetros y criterios de evaluación?. Por ejemplo el número de clases requeridas o la precisión del algoritmo.
   - Al final se debe preguntar si el algoritmo y parámetros elegidos cumplen con el objetivo general del KDD.

3. El paso final es utilizar el algoritmo para buscar patrones y modelos siguiendo los parámetros establecidos, este proceso por lo general está automatizado.

En [Gibert et al., 2010c] se presenta un mapa conceptual con las principales técnicas de minería de datos y varios criterios para apoyar la selección del método más adecuado para usar en cierta aplicación.

2.1.6. Interpretación de los resultados

En esta etapa se interpretan los modelos y patrones hallados en los datos, las técnicas de presentación o visualización son importantes para obtener resultados útiles. *"La calidad de las decisiones tomadas a partir de los procesos de KDD no solo depende de la calidad de los resultados en sí, sino de la calidad del sistema de comunicar dichos resultados de una manera comprensible para la persona que toma las decisiones"*.[Gibert et al., 2008b] La interpretación de los resultados debe ser realizada con ayuda de los expertos en el área de estudio y de acuerdo a los objetivos del análisis. De ser necesario, desde este paso se puede volver a cualquiera de los pasos anteriores para hacer refinaciones o mejoras al proceso.

La presente tesis se acerca a esta linea de interpretación y recoge algunos de los trabajos como CPG[Gibert et al., 2005] o TLP[Gibert et al., 2008a]

2.1.7. Uso del conocimiento descubierto

En último paso consta de documentar el conocimiento adquirido, reportarlo a las partes interesadas e incorporar y usar la información descubierta para la toma de decisiones.

2.2. Introducción a Java-KLASS

Klass es un software que brinda un compendio de herramientas para ayudar a los expertos en el proceso de minería de datos. Klass fue originalmente propuesto y diseñado en la *Facultat d'Informàtica de Barcelona* (FIB) por Karina Gibert como un software orientado a la clasificación automática de dominios poco estructurados en su trabajo de tesis de licenciatura [Gibert, 1991] y luego en su tesis doctoral [Gibert, 1995] en la década de los 90. El software fue en primer lugar desarrollado en LISP y se ejecutaba sobre UNIX, pero luego fue re-escrito en Java puesto que presentaba ventajas como:

- Posibilidad de entregar un ejecutable sin el código fuente. Esto en su versión original presentó problema ya que al ser LISP un lenguaje interpretado no era generar un ejecutable y para usar Klass había que entregar el código fuente.
- Portabilidad y posibilidad de ejecutarse en cualquier sistema operativo.
- Eliminación del costo de las licencias de LISP

Desde su aparición Klass ha ido incorporando nuevas funcionalidades como parte de trabajos de fin de Máster, proyectos del grupo de investigación, tesis Doctorales, o de trabajos en varias asignaturas tanto en los grados de estadística como de ingeniería en informática en la *Universitat Politècnica de Catalunya* (UPC) o la *Universitat Illes Balears* (UIB). Luego de su re codificación en java pasó a Llamarse Java-KLASS y es utilizado hasta la actualidad para proyectos de investigación, docencia y estudios en diferentes ámbitos por lo que es importante con cada actualización o nuevas herramientas que se agregan mantener un registro de las versiones y asegurarse de que las funcionalidades anteriores continúen funcionando.

2.2.1. Funcionalidades de Java-KLASS

En este apartado se presenta un listado de las funcionalidades que ofrece Java-KLASS hasta su más reciente versión.

- Representación de matrices de datos, variables cuantitativas, cualitativas y semánticas y manejo de metadatos.
- Selección de variables e individuos basada en criterios para generar submatrices basado en muestreo aleatorio.
- Recodificación o discretización de variables y generación de variables nuevas.
- Gestión de bases de conocimiento.
- Gestión y visualización de ontologías.
- Gestión y visualización de termómetros.
- Estadística descriptiva extensa univariante, bivariante y trivariante de los datos y de distribuciones condicionadas (CPGs)
- Visualización 3D.
- Análisis dinámico[Gibert et al., 2010a].
- Cálculo de distancias con métricas de distintas familias como: Euclidiana, absoluta, Minkovski, mixta de Gibert [Gibert et al., 2005], ralambondrainy, Gower, Gowda-Diday, Ichino-Yaguchi, mixta de Gibert generalizado, Chi-cuadrada, Hamming generalizado, super concept-base.
- Clusterización automática con métodos jerárquicos clásicos, basados en reglas, en ontologías, con métodos basados en densidades como BDSCAN o OPTICS [Mollá Santiago, 2014] o métodos escalables como CURE.

- Interpretación de clases vía TLP, IRBBP [Gibert et al., 2012b, Gibert et al., 2013] [Gibert and Conti, 2016, Gibert et al., 2008a] y conceptualmente CCEC
- Evaluación de bases de conocimiento *Base de Conocimiento* (BV)
- Métodos de interoperabilidad.
- Gestión de Sistemas heterogéneos que incluye información numérica, cualitativa, semántica, BV, ontologías

2.2.2. Cronología

A continuación se presenta la evolución de Java-KLASS a través del tiempo en la cual se explican las versiones y que incorpora cada una de ellas.

- Feb. 1991 **KLASS v0**. Tesina Karina Gibert. "KLASS. Estudi d'un sistema d'ajuda al tractament estadístic de grans bases de dades". Clasificación de matrices de datos heterogeneas con la distancia mixta de Gibert [Gibert, 1991].
- Nov. 1994 **KLASS v1**. Tesis Karina Gibert. "L'us de la informació simbòlica en l'automatització del tractament estadístic de dominis poc estructurats". Es una ampliación de **KLASS v0** que incorpora la clasificación basada en reglas [Gibert, 1995].
- Jul. 1996 **KLASS v1.1.** PFC Xavier Castillejo. Incorpora a **KLASS.v1** una interfaz de ventanas independiente con un sistema que facilita el uso de KLASS desde SUN y desde PC a usuarios que desconocen Lisp y UNIX. Llamaremos **xcn.KLASS** al núcleo Lisp de esta nueva versión y **xcn.i** en la interfaz C [Castillejo, 1996].
- Oct. 1997 **jj.KLASS.** PFC Juan José Márquez y Juan Carlos Martín. Incorporan a la versión KLASS.v1 nuevas opciones para el tratamiento de datos faltantes, la posibilidad de trabajar con objetos ponderados e implementan un test no paramétrico de comparación de clasificaciones[Márquez and Martın, 1997].
- Sep. 1999 **KLASS v1.2.** PFC Xavier Tubau (versión Beta). Incorpora a la versión **xcn.KLASS** el módulo de comparación de clasificaciones de **jj.KLASS**, la métrica mixta y Ralambondrainy y prepara la formulación de tres más para su posterior implementación. Llamaremos **xt.KLASS** al núcleo Lisp de esta nueva versión y **xt.i** en la interfaz C asociada [Tubau, 1999].
- 1999-2000 **KLASS + v1.** PFC Silvia Bayona. Fusión definitiva de la versión **xt.KLASS** con **jj.KLASS.** Incorpora además un módulo de análisis descriptivo de

los datos, también de las clases resultantes, reorientando KLASS hacia un propósito más general y menos especializado. Llamaremos **sbh.KLASS** al núcleo Lisp de esta nueva versión y **sbh.i** en la interfaz C asociada [Bayona, 2000].

- 2000-2002 **KLASS + v2.** PFC Josep Oliveras. Añade a **sbh.KLASS** las métricas mixtas pendientes (Gower, Gowda-Diday y Ichino-Yaguchi). Llamaremos **joc.KLASS** a esta nueva versión.
- 2000-2003 **jr.KLASS +.** Tesis doctoral Jorge Rodas. Integra **KLASS + v.2** y Columbus, que se introduce más adelante[Rodas-Osollo, 2004]
- 2000-2003 Investigación Anna Salvador y Fernando Vázquez. Desarrollo de **CIADEC**, que se introduce más adelante [Gibert and Salvador, 2000] [Vázquez and Gibert, 2002].
- 2002-2003 **Java-KLASS v0.** PFC Ma. del Mar Colillas. Versión Java del módulo de análisis descriptivo e integración con **CIADEC** y **Columbus**.
- 2003-2005 **Java-KLASS v0.22.** Colaboración con Mar Colillas. Ampliación del análisis descriptivo e introducción de herramientas de gestión de datos (definición de ordenaciones en los informes, posibilidad de varias matrices de objetos en el sistema simultáneamente, cambio de matriz activa).
- 2005-2006 **Java-KLASS v1.0.** Colaboración con Mar Colillas. Incluye la lectura y visualización de dendograma aislados, así como la generación de particiones a partir de ellos.
- 2006-2007 **Java-KLASS v2.0.** PFC Jose Ignacio Mateos. Ampliación de Java-KLASS con un módulo de cálculo de distancias para diferentes tipos de matrices de datos, incluyendo las que combinan información cualitativa y cuantitativa, tratamiento de missing y creación de submatrices.
- 2006-2007 **Java-KLASS v3.0.** PFC Roberto Tuda. Incluye un módulo de clasificación automática por métodos jerárquicos, utilizando todas las distancias implementadas en la v2.0 y una opción para estudiar agregaciones de objetos paso a paso. Se crea la opción de poder seleccionar el directorio de trabajo predeterminado. Se le agrega la opción de añadir y guardar objetos con peso.
- 2006-2007 **Java-KLASS v4.0.** PFC Laia Riera Guerra. Introducción, gestión y evaluación de Bases de Conocimiento. Ampliación de Java-KLASS con un módulo de transformación de variables que permite discretitzacions, recodificación y cálculos aritméticos con variables numéricas. Por último, esta versión incluye la definición de submatrices vía filtros lógicos sobre los objetos, la edición de metainformación de las variables de la matriz, eliminación de variables e importación de archivos en formato .dat estándar.

- 2.007 **Java-KLASS v5.0.** PFC Andreu Raya. Incluye la clasificación condicionada, la clasificación basada en reglas y funcionalidades de división de la base de Datos y de gestión de árboles de clasificación (o dendograma) asociados a las diferentes matrices de datos.
- 2.007 **Java-KLASS v6.0** Trabajo de investigación Tutelada Alejandro García. Clasificación basada en reglas exógenas. Intenacionalització y localización de a tres idiomas (Catalán, Inglés y Castellano). Fusión de matrices.
- 2008 **Java-KLASS v6.4.** Trabajo de Máster Alfonso Bosch Sansa, Patricia García Giménez, Ismael Sayyad Hernando. Boxplot-based discretization, Boxplot-based Induction rules.
- 2008 Tesis doctoral Alejandra Perez. Caracterización por condicionamientos sucesivos, metodología que induce automáticamente a conceptos asociados a las clases descubiertas.
- 2008 Tesis doctoral Gustavo Rodríguez. Clasificación basada en reglas para estados que permite análisis de sistemas dinámicos.
- 2008: **Java-KLASS v7.0.**: TRT Alejandro García Rudolph. Fusión de matrices y gestión de variables activas.
- 2009: **Java-KLASS v8.**: Tesis de máster de Ester Lozano. Criterios Best Local Concept and no close world Assumption del CCECS. PT Alejandro García Rudolph. Clasificación basada en reglas para estados.
- 2010: **Java-KLASS V8.1.**: Práctica SISPD. Narcis Maragall. Boxplot Based Induction Rules
- 2012: **Java-KLASS v8.6.**: Práctica SISPD. Pau. metodología CCECS.
- 2012: **Java-KLASS v9.**: Práctica SISPD. Marco Villegas. Criterios CCECS.
- 2013 **Java-KLASS v10.**: Práctica SISPD. Emili Boronat. Traffic Light Panel.
- 2014: **Java-KLASS v11.**: Proyecto final de Carrera Ingeniería Informática FIB. Sheila Mollà. DBSCAN, OPTICS, 3D Visualization.
- 2014: **Java-KLASS v12.**: Practica SISPD. Jonathan Moreno. Optimización de expresiones lógicas.
- **2015 Java-KLASSv15.**: Practicas IKPDI + SISPD Sergio Santamaria y Daniel Gibert y otros prácticas Gestión de ONTOLOGÍAS, distancias semánticas. Clasificación basada en ontologías.

- 2016 **Java-KLASSv16.**: TFG Valerio Di Matteo (U. La Sapienza, Roma, Italia). Muestreo y Escalabilidad: Generación de variables aleatorias, extracción de muestras aleatorias sobre la matriz de datos, k-Nearest Neighbour, CURE.
- Jun 2016 **Java-KLASSv17.**: TM David Canudes + practicas IKPD des2015: Gestión termómetros + automatización de TLPs.
- Nov 2016 **Java-KLASSv18.**: prácticas IKPD: Implementación de TLPs anotados. Primeras infraestructuras para gestionar variables multivaluadas (desarrollo y concatenaciones)
- Mar 2018 **Java-KLASSv18.2.**: TM Luis Daniel Pérez Tamayo: Gestión de variables multivaluadas y consolidación trabajo anterior.
- May 2018 **Java-KLASSv18.3.**: TM Carlos Luis Jordan y TM Johnny Avila: termómetros cualitativos y conexión con semáforos, y reorganización de todos los métodos de inducción de conceptos.

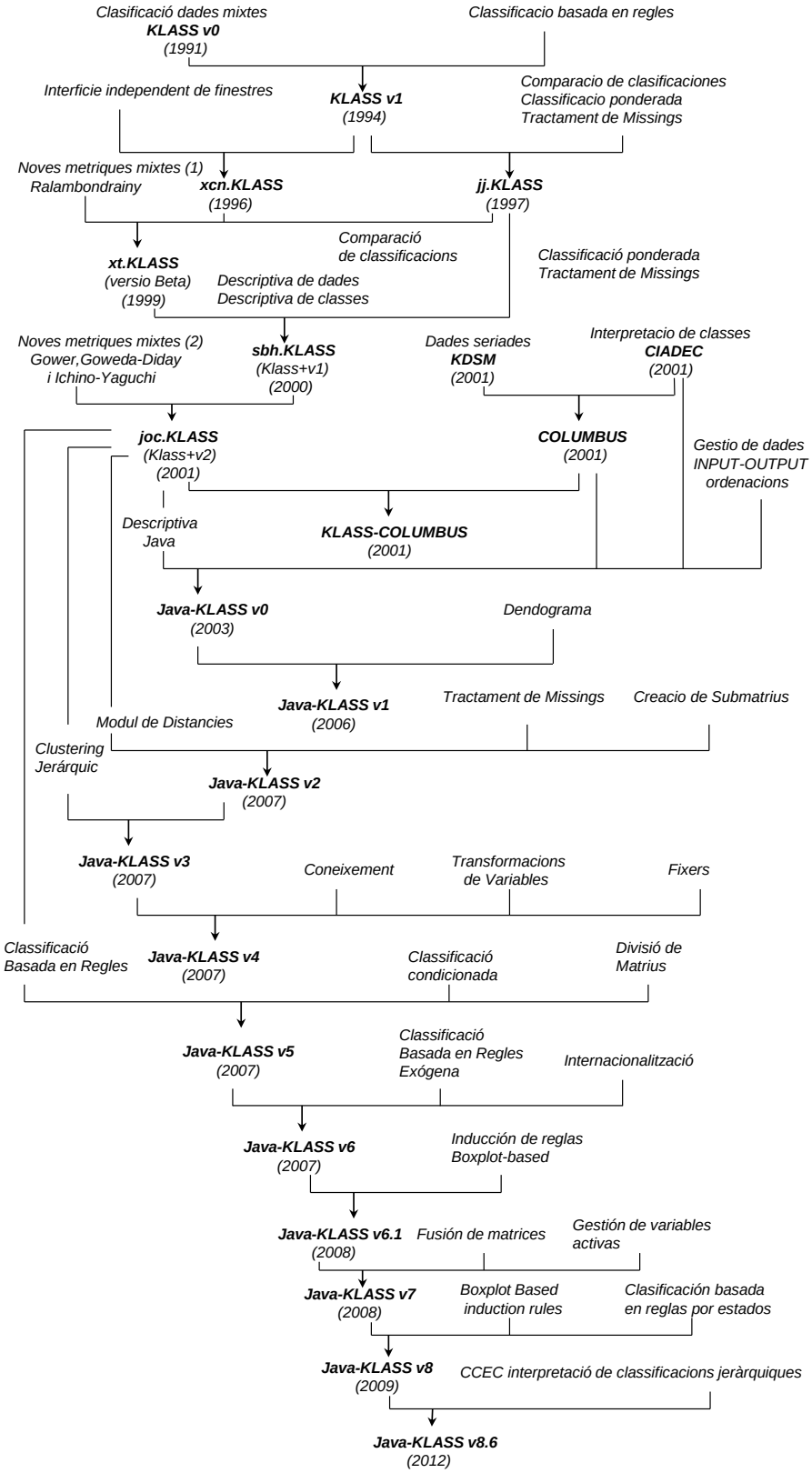

Figura 2.3: *Cronología de Klass. Parte 1*

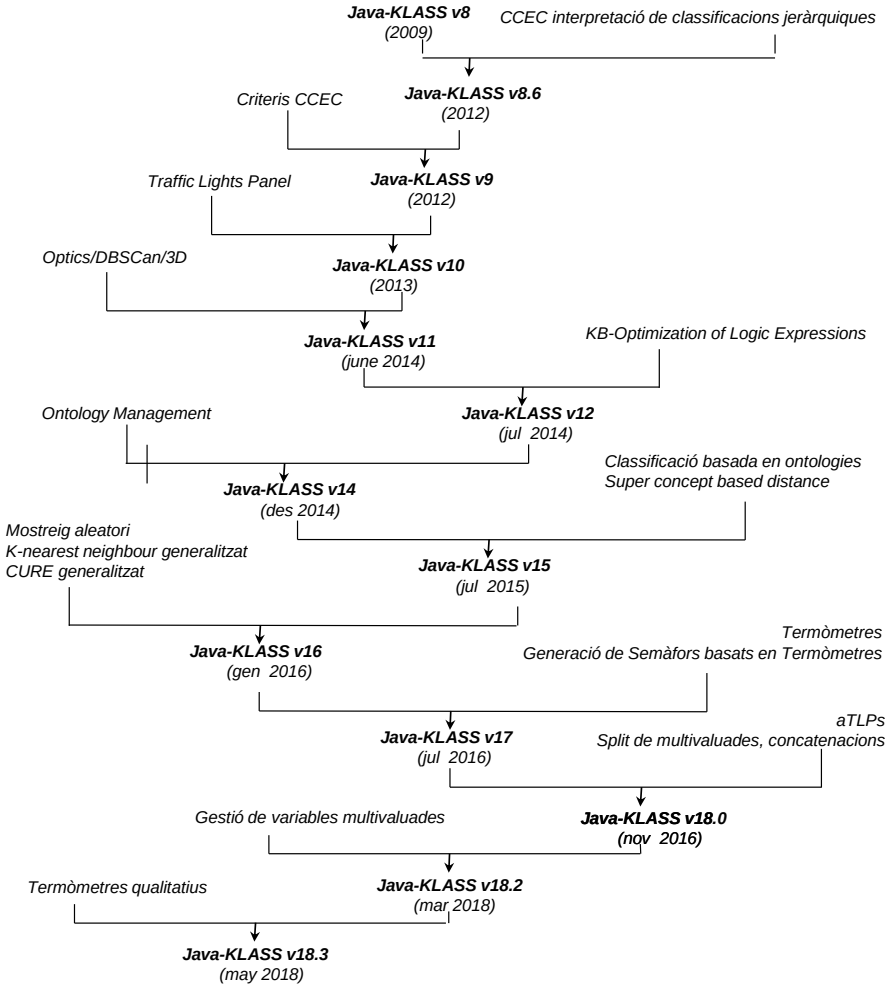

Figura 2.4: *Cronología de Klass. Parte 2*

2.3. Trabajo Realizado

2.3.1. Introducción

En esta sección se presenta una explicación exhaustiva del trabajo realizado para extender los termómetros a variables cualitativas y el proceso para generar el TLP a partir de estas variables. En primer lugar se explica en forma de resumen la conceptualización de las clases, la interpretación mediante semáforos y el funcionamiento de los termómetros con variables numéricas presentado en [Canudes Solans, 2016] y luego se detallan las modificaciones realizadas tanto en la interfaz gráfica, como en las funciones o clases de núcleo. Al final se presentan los semáforos generados a partir de los termómetros extendidos.

2.3.2. Antecedentes

Este trabajo descanza principalmente sobre dos herramientas previas, el TLP [Gibert et al., 2008a] y el termómetro [Canudes Solans, 2016]

Concepción original del TLP

En la sección 1.4.1 se han introducido el CPG como una herramienta de Java-KLASS que ayuda a la conceptualización de las clases resultantes de una clusterización en donde se muestra de forma compacta las distribuciones condicionadas para cada clase y variable. El objetivo del CPG es extraer información útil para interpretar las clases identificando qué variables representan particularidades para cada clase con respecto a las otras[Gibert et al., 2008a]. Sin embargo, la representación basada en histogramas o múltiples boxplots, en aplicaciones reales, puede ser difícil de analizar para expertos sin conocimientos técnicos como médicos, sociólogos entre otros; para esto, se han diseñados los semáforos como una forma más simbólica e intuitiva de representar la información relevante para cada variable y facilitar la interpretación de las clases descubiertas.

El TLP es construido a partir del CPG identificando tres niveles cualitativos a las variables y asignando un color cada para nivel (verde, amarillo y rojo) de acuerdo a la semántica de la variable. El TLP muestra el nivel cualitativo dominante para cada variable dentro de cada clase, el nivel cualitativo dominante se puede encontrar de dos formas: identificando el nivel cualitativo de la media o la

mediana en la clase, o, identificando el nivel cualitativo de la moda en la clase. De esta forma se abstrae la representación basada en histogramas o múltiples boxplots y se presenta la información de forma simbólica que sea de mayor entendimiento para los expertos no técnicos.

Además de esto, en [Canudes Solans, 2016] se introduce un nuevo color (violeta) para representar variables con missings o que no se puedan representar con los 3 colores antes descritos.

Termómetros para variables numéricas

Los termómetros se han diseñado como una herramienta para representar los valores de una variable cuantitativa en un rango de tres zonas de colores de acuerdo al conocimiento previo de los expertos sobre el área de estudio. Se han elegido tres zonas de colores (verde, amarillo y rojo) para mantener relación con los colores del semáforo. Las zonas de colores elegidos sirven para representar la semántica de la variable, el verde representa valores más positivos o benevolentes a la hora de hacer una interpretación, el amarillo valores neutrales y el rojo los valores menos positivos o benevolentes. En la figura 2.5 se muestra el diseño propuesto para el termómetro.

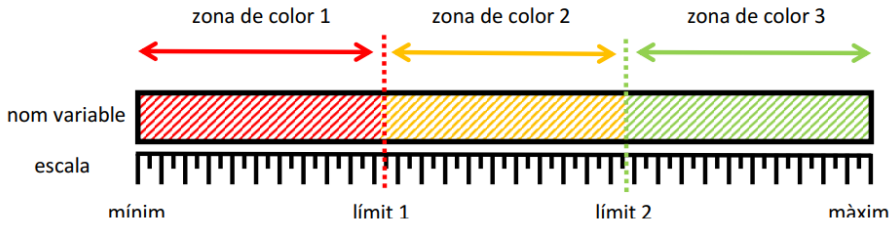

Figura 2.5: *Diseño del termómetro. Fuente: [Canudes Solans, 2016]*

El Diseño de los termómetros se puede re-ordenar de forma que el verde esté mas cercano a los valores mínimos de la variable y el rojo a los valores máximos, de esta forma se puede representar el hecho de que un valor cercano al mínimo en un dato pueda ser bueno o malo a la hora de hacer una interpretación. Como un ejemplo de esto, en [Gibert and Conti, 2015] se hace un estudio de una planta de tratamiento de aguas residuales en el que se presenta una tabla de polaridad semántica en donde para ciertas variables como la concentración de amoniaco en el afluente de la planta piloto, un valor cercano al mínimo es considerado como benevolente, mientras que para otras variables como la temperatura de las aguas

residuales, un valor cercano al máximo es considerado como benevolente.

El objetivo de los termómetros es que a partir de la información de las zonas de color introducida para cada variable se pueda generar automáticamente el TLP por lo que es importante la intervención del experto para definir en el semáforo los puntos de corte en el rango de la variable en donde su semántica cambie para valores buenos, neutrales o malos de acuerdo a lo expuesto anteriormente.

Diseño visual de los termómetros

En la figura 2.6 se puede observar un panel con varios termómetros numéricos. Su diseño ha sido pensado para cumplir con las características expuestas en apartados anteriores. Se puede ver que cada termómetro tiene una barra que representa los valores de cada variable y en la que se pueden establecer los límites o cortes de cada zona de color, dichos límites se pueden ingresar también de forma manual en los cuadros de texto a la derecha del termómetro. Siendo un coloreado directo cuando cuando el rojo está los valores mínimos de la variable. De igual forma se puede invertir los colores del termómetro para que los valores cercanos al mínimo se puedan representar con el verde y los cercanos al máximo con el rojo, en lo que en KLASS se denomina coloreado inverso.

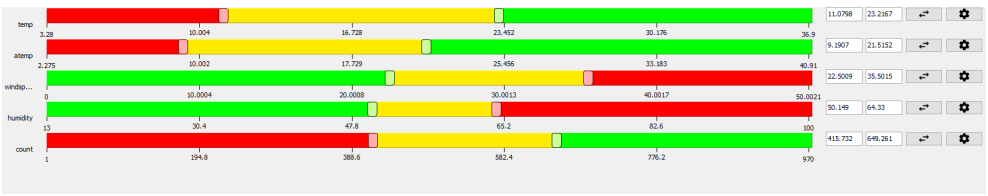

Figura 2.6: *Termómetro en Java-KLASS.*

Estructura original de los termómetros

Para trabajar con el diseño propuesto, se ha pensado en una estructura en forma de matriz de datos en donde cada fila representa un termómetro para cada variable y las columnas representan las características del termómetro. En la tabla 2.2 se muestra la estructura de los termómetros para variables numéricas.

Termómetro						
Núm Termómetro	Propiedades del Termómetro					
	Nom.Var	mínim	limit1	limit2	máxim	ordenación
1	varA	0	10	30	50	R-Y-G
2	varB	0	1000	2000	4000	R-Y-G
...	...	...	...	...	...	...
Ti	**varI**	**0**	**25**	**75**	**120**	**R-Y-G**
...	...	...	...	...	...	...
Tn	VarN	0	0.05	0.25	0.5	G-Y-R

Tabla 2.2: *Estructura de los termómetros numéricos.* ***Fuente****:[Canudes Solans, 2016]*

Como se puede observar, se almacena para cada termómetro, los valores mínimo y máximo de las variables y los 2 cortes para cada zona de color, también se almacena el ordenamiento de los colores. La estructura en forma de matriz de datos es conveniente puesto que se pueden exportar las características de los termómetros a un archivo CSV para que puedan ser recuperados luego por si se necesita volver a trabajar con ellos.

Generación del TLP a partir de los termómetros

Para la generación de los semáforos a partir de los termómetros en primer lugar se debe asociar una variable categórica que generalmente aunque no necesariamente es el resultado de una clasificación o una clusterización, con las variables numéricas que se pudrieron o no haber usado para el proceso de descubrimiento de las clases.

Discretización

A partir del conocimiento previamente ingresado por los expertos en los termómetros, el siguiente paso consiste en discretizar las variables numéricas según las zonas de color ya definidas. para ello se crea una nueva variable cualitativa Z_k que asigna para cada observación o instancia el código del color que le corresponda, según el termómetro, de acuerdo al siguiente método. Que en KLASS se denomina discretización basada en termómetros y está definido para variables numéricas.

Sea X la variable y $\{x_1, x_2, x_3, ..., x_n\}$ el conjunto de valores que pueda tomar dicha variable en una observación específica. Para todo $x_i \in X$, se aplica el algoritmo 1:

Algoritmo 1 Algoritmo de discretización de variables usando termómetros

Entrada: x_i, limit1, limit2, coloreado: variable observada y propiedades del termómetro

Salida: cod: código de color correspondiente (r,y,g,v)

si ($x_i \neq$ NaN) Y ($x_i \in X$) **entonces**
  si ($x_i \leq$ limit1) **entonces**
    si coloreado=*directo* **entonces**
      cod $\Leftarrow$ "r"
    si no
      cod $\Leftarrow$ "g"
    fin si
  si no, **si** ($x_i >$ limit1) Y ($x_i \leq$ limit2) **entonces**
    cod $\Leftarrow$ "y"
  si no, **si** ($x_i >$ limit2) **entonces**
    si coloreado=*directo* **entonces**
      cod $\Leftarrow$ "g"
    si no
      cod $\Leftarrow$ "r"
    fin si
  fin si
si no
  cod $\Leftarrow$ "v"
fin si

Dado el termómetro mostrado en la figura 2.7 para una variable dada, el resultado de aplicar el algoritmo de discretización se puede ver en la tabla 2.3

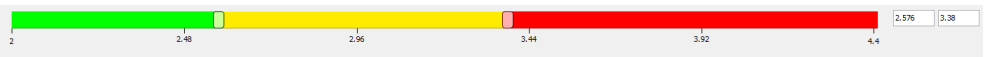

Figura 2.7: *Termómetro en Java-KLASS.*

x_k	Z_k
2.13	g
3.95	r
NaN	v
2.51	g
2.80	y
2.02	g
4.15	g
3.0	y
2.10	g

Tabla 2.3: *Resultado de plicar la discretización sobre el termómetro de la figura 2.7*

Generación de Tablas cruzadas

En el siguiente paso consiste en crear una tabla cruzada en donde las filas representan las modalidades de la variable categórica, que por lo general son las clases descubiertas, y en las columnas los valores de una nueva variable generada que son los códigos de colores que se han obtenido en la discretización de la variable numérica asociada. El número que se coloca en cada celda es el número de objetos de la clase que representan el color de la columna. En la tabla se observa el resultado de crear la tabla 2.4 cruzada para una variable dada.

Variable de Clase	**Variable** Z_k				
Modalidad	R	Y	G	V	Total
C1	90	45	23	0	158
C2	18	72	36	5	131
C3	26	4	61	0	91
C4	81	19	8	4	112
...	...	...	...	...	...
C_i	12	7	32	45	96
...	...	...	...	...	...
C_n	7	51	29	0	87

Tabla 2.4: *Ejemplo de tabla cruzada*

Asignación de colores

Luego de construir la tabla cruzada, se busca el máximo número que aparece en cada fila para asociar al color correspondiente. Esto corresponde a identificar el

color dominante en la clase , que determinará el color de la celda en el semáforo. En caso de empate, el algoritmo original asigna el color de forma aleatoria. En la figura se muestra la asignación de colores a partir de la tabla cruzada $M_k = P \times Z_k$. La tabla $S(P)$ contiene 2 columnas: 1) las modalidades de P (clases) y la segunda el color asignada para una clase $C \in P$, $h(C) = argmax(F_c)$, donde F_c es la fila c de M_k

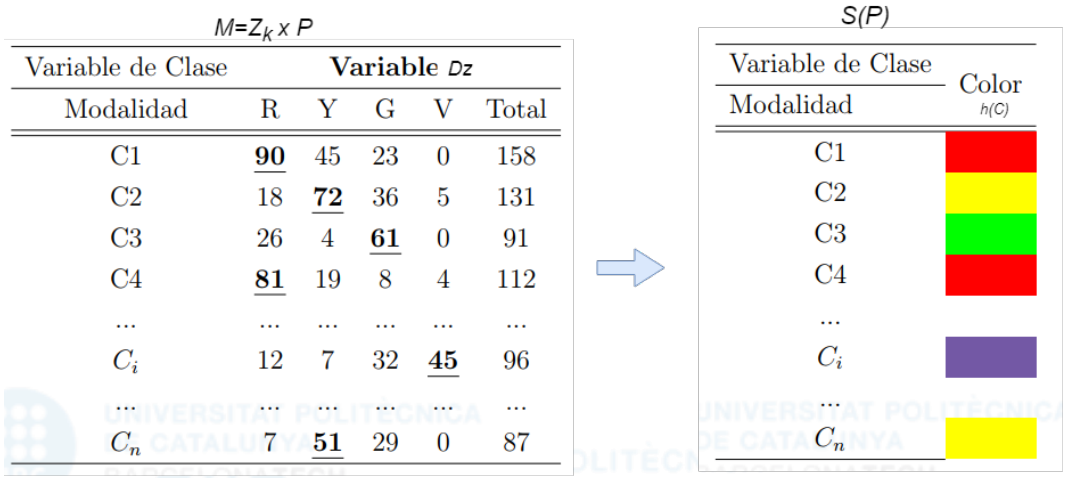

Figura 2.8: *Asociación de colores.*

El resultado de la tabla mostrada en la figura 2.8 representará directamente una columna del TLP, luego se verá que la asignación aleatoria de colores en casos de empate genera ciertas disfunciones que se han corregido en esta tesis con una modificación de este criterio.

Generación del TLP

Como el resultado de aplicar el proceso anterior se tiene la asignación del color para una variable numérica asociada a cada modalidad de la variable categórica, para generar el TLP final, sólo basta con repetir el proceso desde la discretización a la asignación de colores para cada variable numérica que se requiera representar. En la siguiente tabla se puede ver un ejemplo de un TLP completo.

Variable de Clase	Variable numérica asociada							
Modalidad	x_1	x_2	x_3	x_4	x_5	x_5	...	x_n
C1							...	
C2							...	
C3							...	
C4							...	
...	...	...	...	...	...	...	...	...
C_i							...	
...	...	...	...	...	...	...	...	...
C_n							...	

Tabla 2.5: *Ejemplo de TLP*

2.3.3. Desarrollo del termómetro para variables cualitativas

La experiencia de utilizar el TLP basado en termómetros en algunos casos muestra una limitación importante cuando se necesita expresar el significado de una clase en base a variables numéricas y cualitativas simultáneamente. Hasta el momento, la determinación del color de las casillas del TLP en variables cualitativas se ha venido realizando manualmente a partir del análisis del CPG y tablas asociadas. Esta tesis propone extender el termómetro a variables cualitativas como ya se ha dicho.

Estructura propuesta para el termómetro de variables cualitativas

Los termómetros para variables cualitativas no pueden tener el mismo tratamiento que para las cuantitativas, primero porque no tienen un rango continuo de valores sobre los que se pueda establecer límites para las zonas de colores y segundo porque en muchos casos no se puede establecer un orden lógico sobre las modalidades de la variable. Por lo tanto, se trata de idear una forma para dar semántica a las modalidades de cada variable cualitativa, para que luego pueda ser transmitida a los TLP. En la figura 2.9 se muestra el diseño propuesto para el termómetro de variables cualitativas

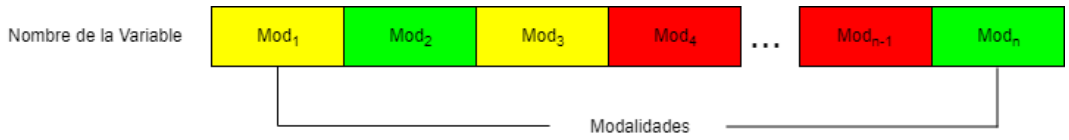

Figura 2.9: *Diseño del termómetro para variables cualitativas.*

Al no existir zonas de corte, se ha diseñado los termómetros para las variables cualitativas de tal forma que todas las modalidades puedan tomar cualquiera de los 3 colores, de esta forma se garantiza que el experto pueda asignar una semántica a cada modalidad por separado pudiendo aparecer en un mismo termómetro más de una vez cada color, así se cumple con la característica que tienen muchas variables cualitativas que no presentan un ordenamiento lógico. Para las variables cualitativas ordinales se puede representar las modalidades con la ordenación correcta. Para estos casos, los colores se conectan por grupos de modalidades.

Propuesta de visualización

En la figura 2.10 se observa en Java-KLASS un panel el termómetro para variables cualitativas. Se han creado siguiendo el diseño antes propuesto, en línea con el diseño para las variables numéricas. El tamaño de las celdas para cada modalidad varía dependiendo del número de modalidades que tenga la variable categórica. De igual forma se han creado los termómetros para que, de acuerdo a la ordenación definida por el usuario en la aplicación, en un mismo panel coexistan los dos tipos de variables, figura 2.11.

Figura 2.10: *Termómetro con variables cualitativas en Java-KLASS*

Figura 2.11: *Termómetro con variables combinadas en Java-KLASS*

Formalización del modelo de termómetros para variables cualitativas

De igual forma que para las numéricas, la estructura de los termómetros para variables cualitativas ha sido pensado en forma de matriz de datos pero con la diferencia de que al no tener un número fijo de puntos de corte, cada fila de la matriz tiene una longitud diferente dependiendo del número de modalidades de la variable. En la tabla 2.6 se muestra la estructura planteada para los termómetros para variables cualitativas.

Termómetro							
N. Ter.	Propiedades del Termómetro						
	Nom. Var.	Ind.	N.Mods	Modalidades			
1	VarQ1	"Q"	3	M1;CodM1	M2;CodM2	M3;CodM3	
2	VarQ2	"Q"	8	M1;CodM1	M2;CodM2	...	M8;CodM8
3	VarQ3	"Q"	4	M1;CodM1	M2;CodM2	...	
...	...	...	...	...	...	...	...
N	VarQn	"Q"	5	M1;CodM1	M2;CodM2	...	M5;CodM5

Tabla 2.6: *Estructura de los termómetros para variables cualitativas*

- En la primera columna se almacena el número de termómetro.
- La segunda columna almacena el nombre de la variable categórica
- La tercera columna contiene la letra "Q" que indica que es una variable categórica
- La cuarta columna almacena el número de modalidades de la variable cualitativa
- Desde la quinta columna en adelante se almacena cada modalidad con su código de color correspondiente

El indicador agregado en la columna 3 se ha planteado para que puedan almacenarse tanto variables categóricas como numéricas en el mismo archivo CSV. En

la figura 2.12 se puede observar un archivo CSV que contiene termómetros para los dos tipos de variable.

```
termo2
NAME;MINIMUM;LIMIT1;LIMIT2;MAXIMUM;COLORING;NUMLABELS;LABELSACCURACY
humidity;13.0;14.74;17.002;100.0;Ascending;5;4
windspeed;0.0;46.95197368240356;48.55204094314575;50.00210189819336;Descending;5;4
atemp;2.2750000953674316;38.82370986080169;40.36910985088348;40.90999984741211;Ascending;5;4
motivation;Q;14;Sopar;r;Cafe;r;Teatre;r;Desconegut;r;Vermut;r;Treballa;r;Encarrecs;r;Compres;r;
WeatherRec;Q;3;Clouds;g;LightRain;g;Acceptable;g;
```

Figura 2.12: *Archivo CSV con termómetros para variables cuantitativas y cualitativas*

Objeto en java de los termómetros

Partiendo de los modelos de termómetros antes propuestos, se han diseñado los objetos en java y se realizado una modificación al modelo de clases original descrita en [Canudes Solans, 2016] que se presenta en la figura 2.13.

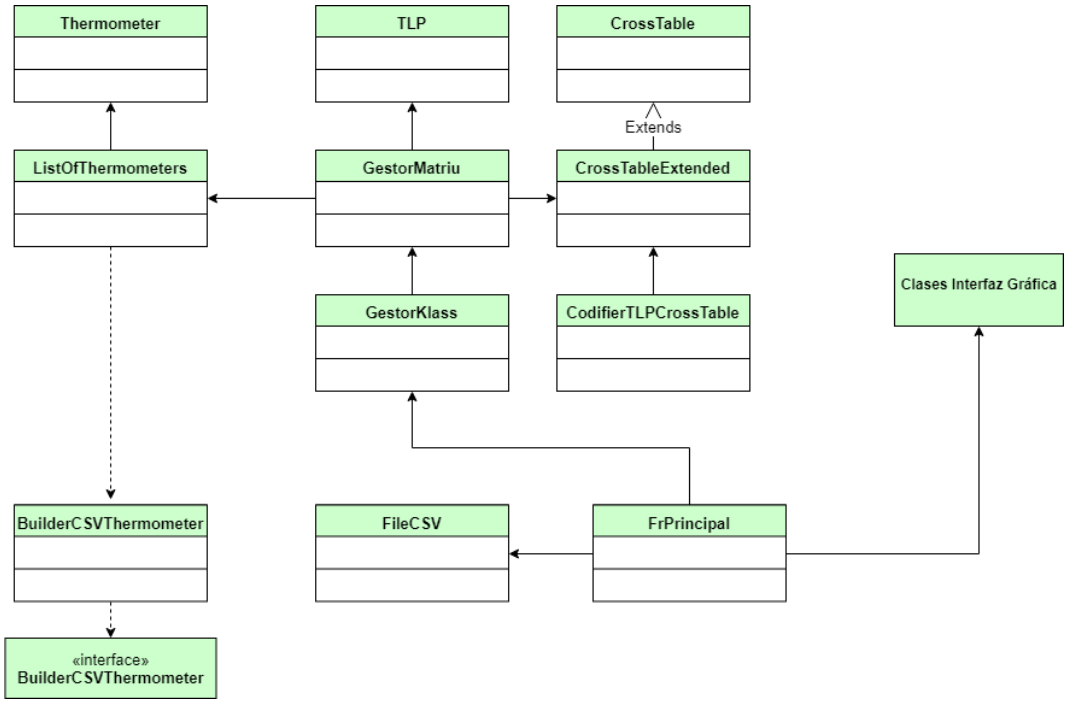

Figura 2.13: *Diseño original del diagrama de clases de los termómetros*

Con base en este modelo, se ha generalizado la clase *"Thermometer"* y se han creado las clases *"ThermometerN"* y "ThermometerQ" que heredan de ella para introducir los termómetros para variables cualitativas al diseño original, con esta generalización se asegura que la funcionalidad original siga intacta y que no se tengan que hacer mayores cambios en el modelo, para esto, el sistema verifica con

que tipo de termómetro está trabajando y actuá acorde a la funcionalidad de cada uno. En la figura 2.14 se muestra el diagrama de clases modificado.

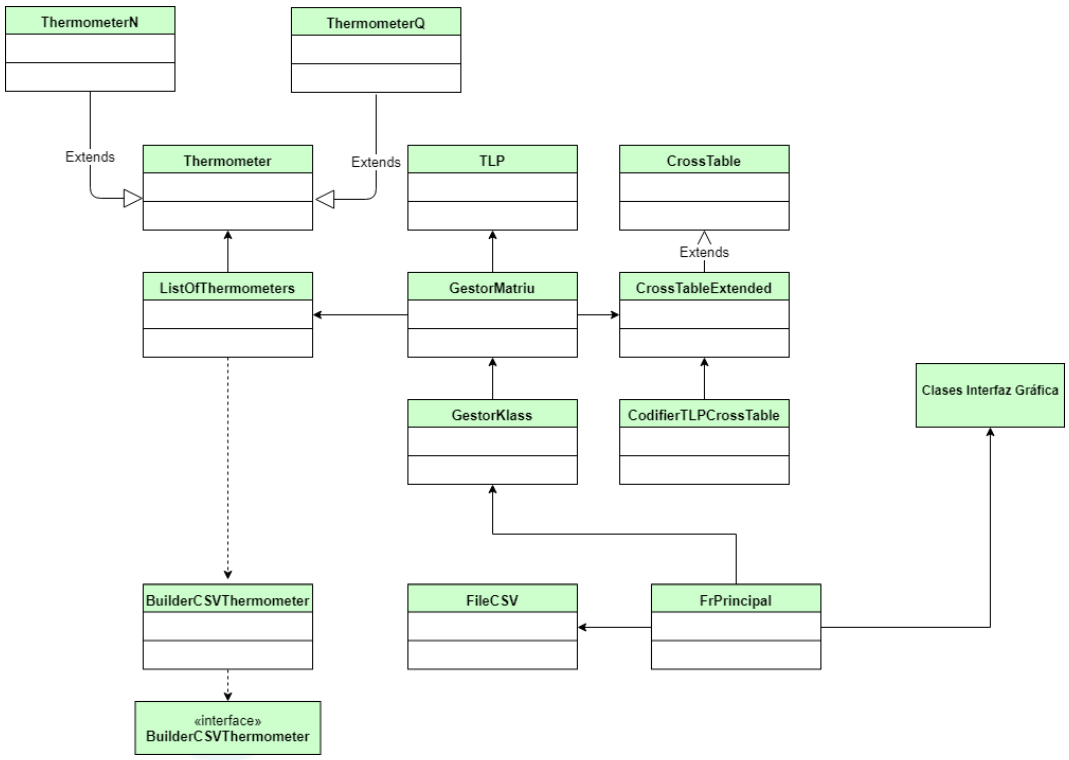

Figura 2.14: *Diseño original del diagrama de clases de los termómetros*

La clase *"ListOfThermometers"* contiene una lista con un termómetros para cada variable, con la generalización se logra que la lista pueda contener termómetros para variables cualitativas y cuantitativas. Además, la clase antes mencionada está relacionada con el gestor de la matriz y ésta con el gestor de Klass, haciendo que los termómetros sean accesibles desde cualquier parte de la aplicación.

La clases *"FileCSV"* sirve para exportar los termómetros a un archivo CSV, mientras que la clase y *"BuilderCSVThermometer"* sirve para importar termómetros desde un archivo CSV, de igual forma se han modificado las funciones de estas clases para que sea posible exportar e importar los dos tipos de termómetro.

2.3.4. Generación del TLP a partir de los termómetros para variables cualitativas

El proceso de generación automática del TLP a partir de los termómetros para variables cuantitativas que se ha descrito en 2.3.2 debe ser modificado puesto que la asignación de colores para las casillas del TLP se realiza en base a la discretización de la variable de acuerdo al su rango de valores y a las zonas de colores definidas en los termómetros. Al no poder realizar la discretización en variables cualitativas, se debe buscar una nueva forma de asignar un color a las casillas, para esto se ha utilizado el proceso de recodificación.

Recodificación

Al igual que en el proceso anterior, el primer paso es obtener un código de color para la variable, pero en lugar de usar la discretización en donde se obtiene el código deseado de acuerdo al valor numérico de la observación y las zonas de color, para las variables categóricas se usa la recodificación, el objetivo de este proceso es obtener un código de color de acuerdo al color asignado manualmente a la modalidad de la variable cualitativa, esta operación se parece a un mapeo directo y se explica a continuación.

Sea X_k una variable categórica y D_k el conjunto de posibles valores k que puede tomar x_k, para todo $m_m \in D_k$ se aplica el algoritmo 2.

Algoritmo 2 Algoritmo de recodificación de variables usando termómetros

Entrada: X_k, $\{Mod1; codMod1, Mod2; codMod2, ..., Modn; codModn\}$: variable a tratar y modalidades con el código de color ingresadas en el termómetro.

Salida: cod: código de color correspondiente (r,y,g,v)

1: encontrado $\Leftarrow FALSE$
2: **para todo** $(Mod; codMod \in \{Mod1; codMod1, Mod2; codMod2, ..., Modn; codModn\})$ **hacer**
3: **si** $(Mod{=}q_i)$ **entonces**
4: cod $\Leftarrow$codMod
5: encontrado $\Leftarrow TRUE$
6: **fin si**
7: **fin para**
8: **si** encontrado $\neq TRUE$ **entonces**
9: cod $\Leftarrow$ "v"
10: **fin si**

En el caso de las variables cualitativas se establece una relación modalidad - color de forma que la variable Z_k, que toma los valores en $Dz = \{r, g, y, v\}$, es una agregación de los valores de X_k y por la tanto la tabla $M_k = P \times Z_k$ es una agregación se las columnas de la tabla $P \times Z_k$. Después de la recodificación se crean las tablas cruzadas siguiendo el mismo proceso explicado en la sección 2.3.2.

Asignación de Colores

El proceso de asignación de colores después de la recodificación es similar al explicado en la sección 2.3.2, sin embargo, se hacen una leve modificación para distinguir entre cualitativas binarias y no binarias.

- **Variables Cualitativas binarias:** El término binarias se usa para hacer referencia a las variables cualitativas que tienen únicamente dos modalidades, por ejemplo, una variable de este tipo puede expresar la presencia o la ausencia de cierta proteína en la sangre en un test médico. El problema aquí es que al poder asignar solo un color por cada modalidad, no es posible generar directamente un semáforo para esta variable con los tres colores que han descrito. Esto puede representar un problema en casos prácticos, como se verá mas adelante, cuando al construir un TLP manualmente puede ser que el experto asigne 3 colores para una variable binaria. En este tipo de variables se presupone que por lo general la una modalidad de la variable puede ser expresada por el experto como benévola y se le asigna el color verde, mierras que la otra modalidad como malévola y se le asigna el color rojo, por lo que no existe el color amarillo o neutral.

  Para que sea posible generar automáticamente un semáforo con tres colores a partir de un termómetro para una variable binaria se ha introducido una nuevo término llamado gamma (γ) y para la asignación del color se usa el proceso indicado a continuación:

  Dada una fila de la tabla cruzada descrita en 2.3.2, en donde $numR$ representa el número de rojos, $numG$ el número de verdes, $numV$el número de violetas o *missings*, y *total* la suma de verdes, rojos y violetas, como se ha dicho antes, el número de amarillos debería ser igual a cero (0). El color para el semáforo es asignado de acuerdo al algoritmo 3.

Algoritmo 3 Algoritmo asignación de color para variables cualitativas binarias

Entrada: $numR$, $numV$, $total$, γ: fila de la tabla cruzada y gama.

Salida: *color*: color correspondiente (r,y,g,v)

1: **si** ($numV < numR$) Y ($numV < numG$) **entonces**
2: **si** ($numR > numG$) **entonces**
3: **si** (numR/total $\geq \gamma$) **entonces**
4: color $\Leftarrow$ "r"
5: **si no**
6: color $\Leftarrow$ "y"
7: **fin si**
8: **si no**, **si** ($numR > numG$) **entonces**
9: **si** (numG/total $\geq \gamma$) **entonces**
10: color $\Leftarrow$ "g"
11: **si no**
12: color $\Leftarrow$ "y"
13: **fin si**
14: **si no**
15: color $\Leftarrow$ "y"
16: **fin si**
17: **si no**
18: color $\Leftarrow$ "v"
19: **fin si**

Un ejemplo de la aplicación del proceso para asignación del color explicado anteriormente se muestra en la tabla 2.7, para realizar el ejemplo se ha asignado un valor de $\gamma = 0{,}6$

Variable de clase P	Tabla cruzada para la variable binaria					
	r	**y**	**g**	**v**	**total**	**Color**
C1	15	0	12	0	27	
C2	7	0	6	1	11	
C3	4	0	16	0	20	
C4	18	0	16	0	34	
C5	19	0	3	4	26	
C6	4	0	3	10	17	
C7	2	0	18	0	20	

Tabla 2.7: *Aplicación del algoritmo de asignación de color con un valor de* $\gamma = 0{,}6$

- **Variables cualitativas no binarias:** la asignación de colores para las variables cualitativas con más de dos modalidades se hace igual que en el caso de las variables numéricas.

Gestión de empates

Otra modificación que se ha realizado al algoritmo de generación del TLP a partir de termómetros original es que en caso de existir empate entre algún color con el amarillo o entre los tres colores en la tabla cruzada explicada en la sección 2.3.2, no se asignará un color de forma aleatoria sino se opta por un modelo más conservador y se da preferencia al color amarillo La opción anterior introducía un cierto indeterminismo en el TLP que para aplicaciones reales no era bien recibido. Al eliminar la aleatoriedad, se asegura que la generación del TLP siempre reproducible el número de veces que se requiera.

Formalización:

Sea un TLP en donde las filas $C \in P$ representan una clase de la variable categórica P y X_k la variable asociada. Sea Z_k el resultado de discretizar o recodificar la variable X_k de acuerdo a su termómetro, y sea $F_c \in M_k$ una fila de la matriz cruzada $M_k = P \times Z_k$, $F_c = \{n_{cr}, n_{cg}, n_{cy}, n_{cv}\}$ donde $n_{cr}, n_{cg}, n_{cy}, n_{cv}$ son las veces que aparece el color rojo, verde amarillo o violeta respectivamente en la clase C representando a esta fila de la matiz. El color de la celda es denotado como S_c y es expresado de la siguiente forma:

Dado una clase $C \in P, S_c$ se modifica respecto a su definición original si F_c

presenta empate.

- Si $(card[argmax(F_c)] > 1) \wedge (argmax(F_c) = 3)$ entonces $S_C = y$

En donde "3" es el subíndice de la columna del amarillo.

Flexibilización de la Generación del TLP basada en termómetros

La generación automática del TLP propuesta en [Canudes Solans, 2016] se ha diseñado para que funcione si y solo si existe previamente un termómetro para todas las variables numéricas asociadas a la variable categórica o de clase que se quieran representar con los semáforos. En este trabajo se propone un cambio a este proceso y se flexibiliza la generación de los semáforos para que se aproveche el conocimiento previo que puede existir para las variables numéricas y cualitativas y se pueda generar el TLP con los termómetros existentes y con variables que no tenga un termómetro asociado. Para estas últimas se crea una fila completa de celdas amarillas para que el experto pueda editarla luego según su criterio en el post proceso de los resultados.

2.3.5. Fomalización del proceso generalizado

El proceso final de creación del TLP se explica a continuación:

Sean $\{X_1, X_2, X_3, ..., X_K\}$ las variables cualitativas o cuantitativas que se quieren representar en un TLP, $T = \{t_1, t_2, t_3, ..., t_K\}$ el panel de termómetro disponible, en donde $t_k, k \in \{1 : K\}$ es el termómetro para la variable X_k, y sea P una variable categórica que actúa como variable de clase para la construcción del TLP. La generación del cuadro semáforo se compone del vector $S = \{s_1, s_2, s_3, ..., s_K\}$ en donde S_k es el semáforo basado en t_k. La construcción es la siguiente:

1. Fase de Discretización/Recodificación
   - Si X_k es númerica y $t_k = r_1, r_2, o \in T$ entonces crear $Z_k = dis(X_k, t_k)$
   - Si X_k es categórica y $t_k = \{(m_1; q_1), (m_2, q_2), ..., (m_{nk}, q_{nk})\} \in T$ entonces crear $Z_k = rec(X_k, t_k)$
   - Si $t_k \notin T$ entonces crear $Z_k = \{y, y, y, ..., y\}$

En donde $nk = card(D_k)$, Z_k es una nueva variable resultante de la recodificación o discretización de X_k según su tipo original, y $dis(X_k, tk)$ y $rec(X_k, tk)$ son las operaciones de discretización, recodificación. Si el panel de termómetro no contiene un termómetro para la variable X_k, se aplica amarillo a todas las clases y el usuario lo editará manualmente. Estos procesos se describen a continuación.

- **Discretización**:
  Dado $\{x_1, x_2, x_3, ..., x_N\}$ el conjunto de valores que puede tomar la variable numérica X_k para un conjunto de N individuos, y $t_k = \{r_1, r_2, o\}$ el termómetro para la variable X_k en donde r_1 y r_2 son los puntos de corte del termómetro y o es la polaridad semántica de la variable ($o \in \{directa, inversa\}$). Para todo $x_n, n \in [1 : N]$, se define z_n como la observación discretizada que obtiene de la siguiente manera:
  - Si $x_n \in X_k$
    - Si $x_n \leq r_1$
      - Si $o = directa$ entonces $z_n =$"r"
      - Si $o = inversa$ entonces $z_n =$"g"
    - Si $(x_n > r_1) \wedge (x_n \leq r_2) \rightarrow z_n =$"y"
    - Si $x_n > r_2$
      - Si $o = directa$ entonces $z_n =$"g"
      - Si $o = inversa$ entonces $z_n =$"r"
  - Si x_n es missing entonces $z_n =$"v"
- **Recodificación:**
  Dado $\{x_1, x_2, x_3, ..., x_N\}$ el conjunto de modalidades que puede tomar la variable categórica X_k y $t_k = \{(m_m; q_m), \forall m \in D_k\}$: $q \in D_z$ en donde (m_m, q_m) es el par modalidad-color que se almacena en el termómetro para variables cualitativas. Para todo $x_n, n \in [1 : N]$, se define z_n a la observación recodificada que obtiene de la siguiente manera:
  Siendo $t_k = \{m_m, q_m\}_{m=1:nk}$
  - Si $(x_n = m_m)$ entonces $Z_k = q_m$
  - Si x_n es missing entonces $Z_k =$ "v"
- **Asignación de amarillo:**
  Dado $\{x_1, x_2, x_3, ..., x_N\}$ el conjunto de valores que puede tomar la variable numérica o categórica. Para todo $x_n, n \in [1 : N]$, se define z_n a la

observación asignada a amarillo que obtiene de la siguiente manera: $z_n =$ "y"

2. Fase de creación de la matriz cruzada

$$M_k = P \times Z = \begin{bmatrix} n_{11} & n_{12} & n_{13} & n_{14} \\ n_{21} & n_{22} & n_{23} & n_{24} \\ n_{31} & n_{32} & n_{33} & n_{34} \\ \vdots & & & \\ n_{c1} & \dots & n_{cq} & \dots \end{bmatrix}$$

En donde $n_{cq}, c \in P$ representa el número de veces que aparece el código de color $q \in D_z = \{r, g, y, v\}$ para la clase C. Y $n_C = \sum_{q=1}^{4} n_{cq}$, $N = \sum_{\forall c \in P} n_c$.

3. Fase de asignación de color Sea $F_c \in M_k$ una fila de la matriz anterior, El color de la celda es denotado como S_C y es expresado de la siguiente forma:

   - **Variables cualitativas binarias**
     - Si $argmax(F_c) = 4$ entonces $S_C = v$
     - Si $(argmax(F_c) = 3)$ entonces $S_C = y$
     - Si $argmax(F_c) = 1$
       - Si $n_{c1}/n_C \geq \gamma$ entonces $S_C = r$
       - Si $n_{c1}/n_C < \gamma$ entonces $S_C = y$
     - Si $argmax(F_c) = 2$
       - Si $n_{c2}/n_C \geq \gamma$ entonces $S_C = g$
       - Si $n_{c2}/n_C < \gamma$ entonces $S_C = y$

     En donde r,g,y,v son los colores rojo, verde, amarillo y violeta respectivamente.

   - **Resto de variables**
     - Si $card[argmax(F_c)] = 1$ entonces $S_C = q_{argmax(F_c)}$
     - Si $card[argmax(F_c)] > 1 \wedge (argmax(F_c) = 3)$ entonces $S_C = y$

     En donde $q_{argmax(F_c)}$ es el color asignado de acuerdo a la posición que ocupa q en $Dz = \{r, g, y, v\}$

Capítulo 3

Caso de Estudio

3.1. Introducción

En este capítulo se usan los semáforos de variables categóricas y numéricas para la generación automática del TLP y la interpretación de las clases resultantes al aplicar un proceso de *Profiling* a los sistemas de salud mental de países en vías de desarrollo. Para esto, se utilizan los datos brindados por la OMS que es la entidad directora y coordinadora de la salud de dentro del Sistema de las Naciones Unidas.

Al decir sistemas de salud mental se hace referencia a las"*Estructuras y todas las actividades cuyo proposito primario es promover, mantener o restaurar la salud mental. Los sistemas de salud mental incluyen a todas las organizaciones y recursos que se enfocan en mejorar la salud mental*". Los datos utilizados provienen de un estudio realizado en países de bajos o medios ingresos, o en inglés *Low and middle income countries* (LAMIC), con el objetivo de probar si una mayor depresión en madres jóvenes en estos países es una causa importante de la mortalidad de los neonatos.

En primer lugar se describe los datos utilizados, luego se modelan los semáforos, se realiza la generación automática del TLP y se presentan los resultados obtenidos con la generalización de termómetros realizada en éste proyecto.

3.2. Descripción de los datos de la OMS

La base de datos denominada *"WHO-AIMS v2.2"* compila un conjunto de 42 países clasificados como LAMIC seleccionados entre febrero de 2005 y febrero de

2008 está compuesta por 22 facetas, 155 ítems y 256 variables y tiene en cuenta los siguientes dominios:

- La política y el marco legislativo del país.
- Los servicios de salud mental.
- La salud mental en la atención primaria.
- Los recursos humanos.
- La información pública y los vínculos con otros sectores.
- La monitoreo e investigación.

Además de esto cuenta con indicadores compuestos adicionales dados por la OMS y conocimiento previo de los expertos.

3.3. Clasificación o *profiling* de los sistemas de salud mental

En [Gibert et al., 2010b] se describe un evaluación de la base de datos "WHO-AIMS" de acuerdo a la siguiente metodología:

1. En primer lugar se hace un análisis del dataset utilizado similar al que se realiza en la sección anterior.
2. Despues, con ayuda de los expertos se selecciona un primer conjunto de variables que contienen la información característica de los 6 dominios descritos anteriormente, y que además están relacionadas con los indicadores compuestos (variables de decisión) de la OMS y con el conocimiento previo de los expertos.
3. En el siguiente paso, con las variables seleccionadas, se realiza la imputación de los datos faltantes o missings utilizando el método MIMMI, este proceso se encuentra ampliamente documentado en [Gibert, 2014].
4. A continuación se introducen variables que representan el nivel de ingreso de los países según los datos de la clasificación del banco mundial, la región geográfica a la que pertenece cada país y el conocimiento previo de los expertos para generar un conjunto de reglas.

5. A continuación se realiza un análisis cluster basado en reglas utilizando el criterio de Ward [Murtagh and Legendre, 2014] y las métricas mixtas de Gibert [Gibert and Cortés, 1997].

6. Como parte de este proyecto se han reproducido los resultados de [Gibert et al., 2010b] usando la nueva versión de Java-KLASS. Como resultado de este proceso se obtiene el árbol mostrado en la figura 3.1 el cual ha sido tallado en 7 clases.

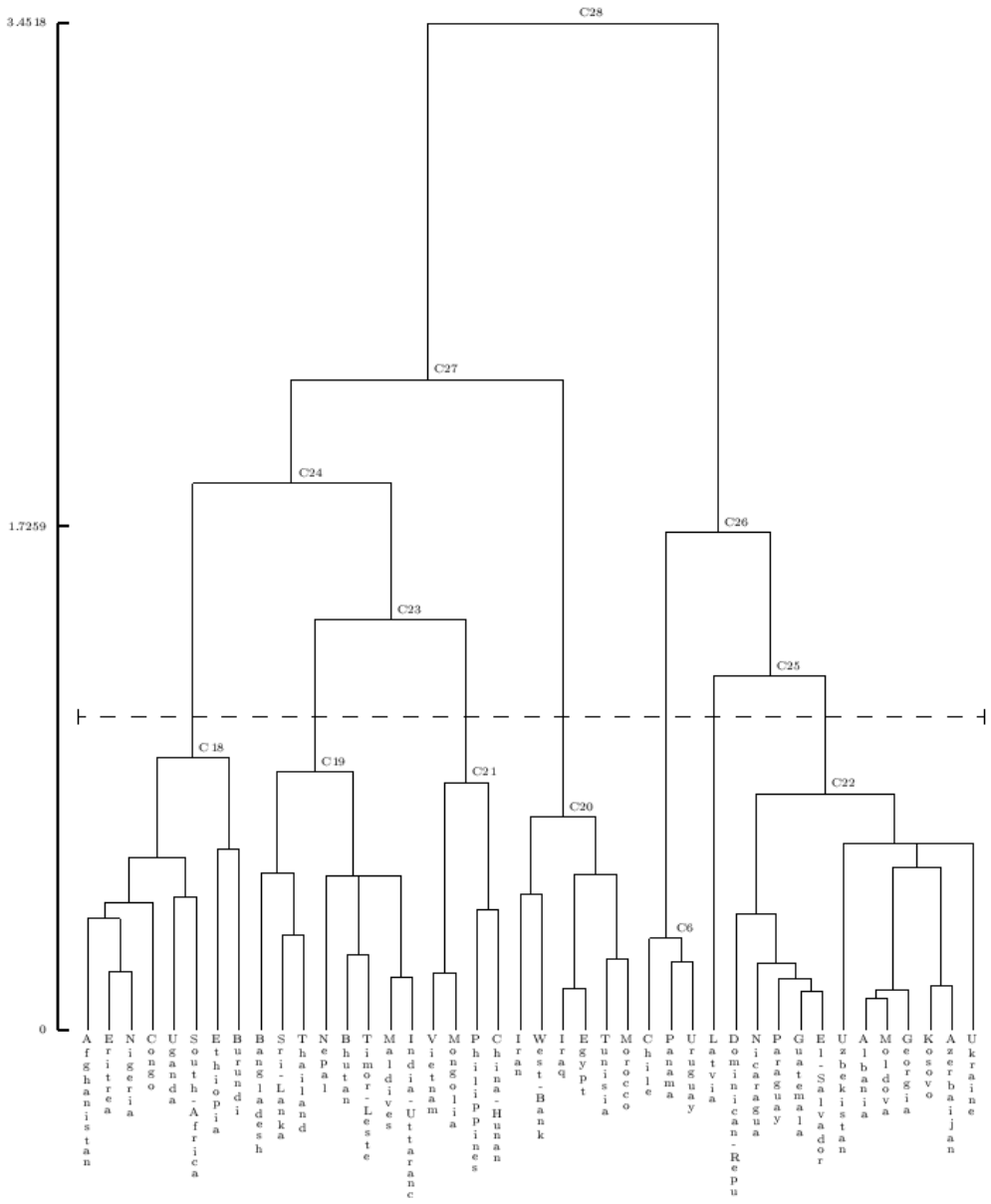

Figura 3.1: *CAJ: Árbol general de clasificación tallado en 7 clases.*

A continuación en la tabla 3.1, se muestran las clases descubiertas y los países

pertenecientes a cada clase y en la figura 3.2 se muestra un mapa de los países pintados con el color de su clase correspondiente.

Clase	Objetos
C18	Afghanistan, Burundi, Congo, Eritrea, Ethiopia, Nigeria, South-Africa, Uganda
C22	Albania, Azerbaijan, Dominican-Repu, El Salvador, Georgia, Guatemala, Kosovo, Moldova, Nicaragua, Paraguay, Ukraine, Uzbekistán
C19	Bangladesh, Bhutan, India - Uttarakhand, Maldives, Nepal, Sri-Lanka, Thailand, Timor -Leste
C6	Chile, Panama, Uruguay
C21	China-Hunan, Mongolia, Philippines, Vietnam
C20	Egypt, Iran, Iraq, Morroco, Tunisia, West-Bank
Latvia	Latvia

Tabla 3.1: *Clasificación de los servicios de salud mental en los distintos países*

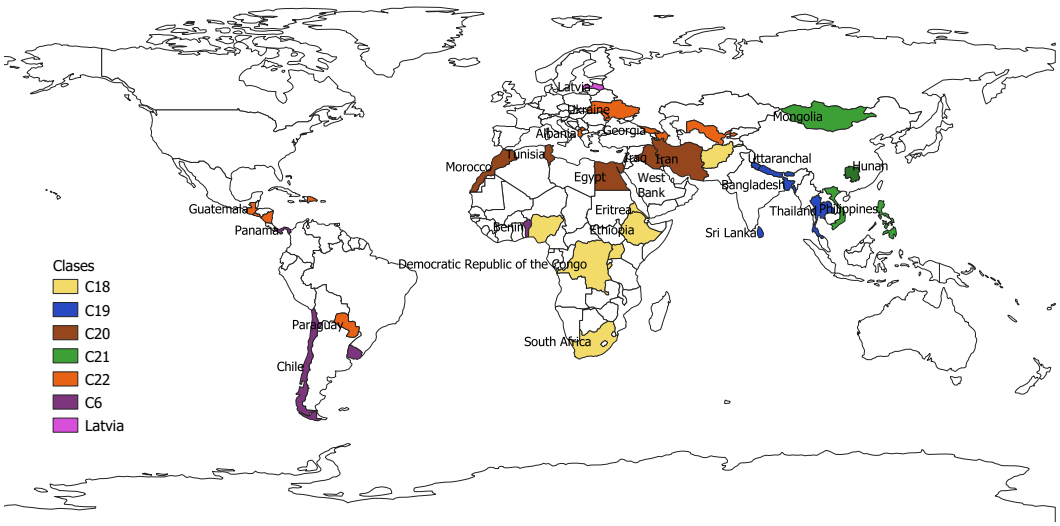

Figura 3.2: *Mapa de los países con su clase correspondiente.*

7. El siguiente paso es la interpretación de las clases, para esto, con ayuda de los expertos se han elegido 14 variables relevantes para la toma de decisiones, en la tabla 3.2 se describen las variables seleccionadas.

Variable	Significado	Tipo de Variable
Incgroup	Nivel de ingreso del país.	Cualitativa
totprofmh	Número total de profesionales dedicados a la salud mental en el país por cada 10000 habitantes.	Cuantitativa
usmhexperca	Gasto en salud mental per cápita en USD.	Cuantitativa
treatpre	Parte de la población diagnosticada y atendida por cada 100000 habitantes.	Cuantitativa
capratiosch	Cobertura del tratamiento de la esquizofrenia.	Cuantitativa
d2f11i1closepsybeds	Camas psiquiátricas ubicadas en o cerca de la ciudad más grande (proporción per cápita).	Cuantitativa
d1f5i2exmhos	Gasto en hospitales mentales. (%)	Cuantitativa
d2f6i71mhrec10y	Proporción de pacientes que permanecen en hospitales psiquiátricos durante 10 años o más.	Cuantitativa
comcarewor	Proporción de usuarios tratados en hospitales mentales.	Cuantitativa
lundpararectrail	Ratio entre consultas externas y días en que el paciente está hospitalizado, indica si el sistema de salud da prioridad a mantener al paciente o ingresarlo lo más pronto posible.	Cuantitativa
D3f1i3Manuals	Disponibilidad de manuales de tratamiento y evaluación en atención primaria.	Cualitativa
Legisl	Presencia de legislación.	Cualitativa - Binaria
Polplanr	Presencia de un plan de salud menta.l	Cualitativa - Binaria
d6f1i6govmhrep	Informe sobre salud mental publicado por el departamento de salud del gobierno.	Cuantitativa

Tabla 3.2: *Descripción de las variables usadas para la interpretación*

Con las variables descritas anteriormente, el siguiente paso es crear el CPG que se muestra en la figura 3.3. A partir del gráfico mencionado, para realizar la interpretación,se analiza el CPG y se construye el TLP, para lo cual se asigna el color verde a los valores de las variables que indiquen sistemas de salud más desarrollados, en la figura 3.4 se muestra el TLP resultante creado manualmente con ayuda de los expertos.

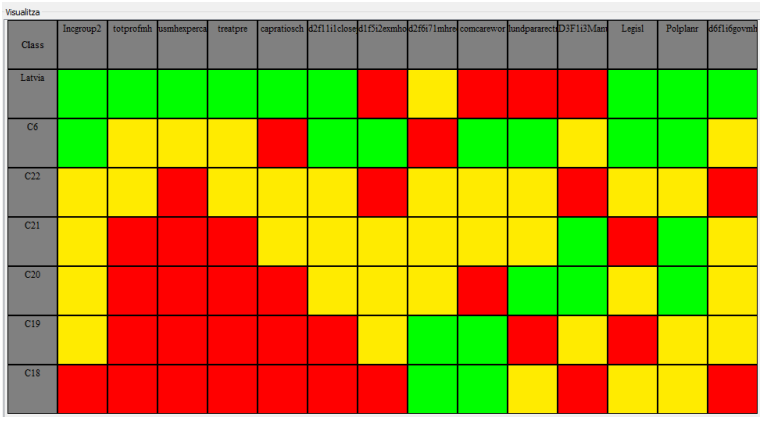

Figura 3.4: *TLP creado manualmente con ayuda de los expertos*

Como resultado de este proceso, se obtiene los 7 perfiles de sistemas de salud mental que se describen de la siguiente forma (Descripciones tomadas literalmente de [Gibert et al., 2010b]):

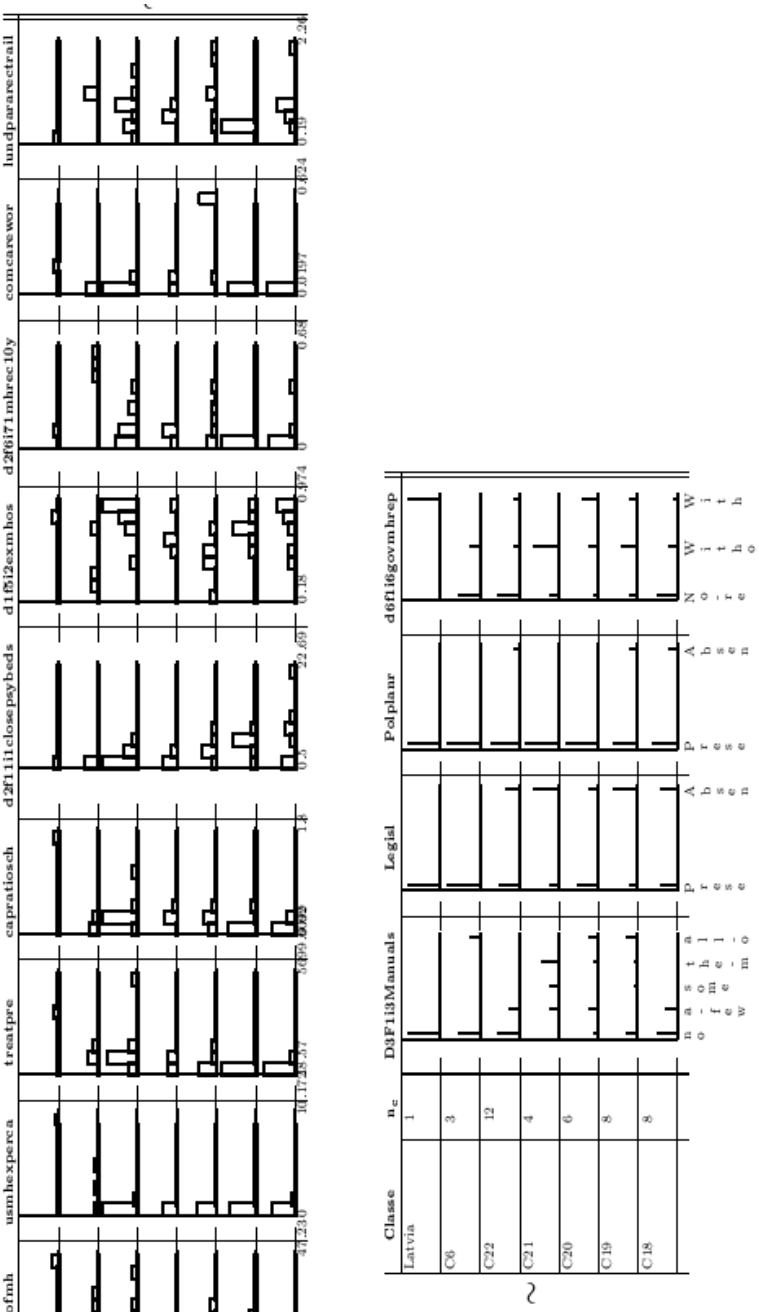

Figura 3.3: CPG con las clases descubiertas y las variables para la interpretación.

Clase Latvia (Latvia) *"En esta clase se observa un alto de ingresos segun la variable "Incgroup", un alto gasto en salud mental y alto número de recursos humanos y materiales, aunque un sistema altamenete dirigido hacia la atención hospitalaria de acuerdo a la variable "lundpararectrail". Es un sistema altamente descentralizado (las camas para pacientes hospitalizados se distribuyen en el país, valores bajos en la variable "d2f11i1closepsybeds"), con una buena covertura (la cobertura de trastornos esquizofrénicos es la más alta), pero se basa en gran medida en la hospitalización de los pacientes en hospitales psiquiatricos. El componente de atención primaria de salud probablemente no está bien desarrollado. Los datos del sistema de información se difunden y están disponibles en el informe del departamento de salud mental. Según la opinión del experto, parece que este sistema de salud tiene algunos problemas de eficiencia con respecto a los gastos excesivos en la atención hospitalaria".* De aquí, se define a la clase como: ***"Sistema mental rico pero basado completamente en hospitales mentales."***

C6 (Chile, Panama,Uruguay): *"Estos países se encuentran en el grupo de ingreso medio alto. La disponibilidad de recursos no es tan alta como en "Latvia", sin embargo están bien equipados, con mayores gastos en salud mental per cápita que otros grupos. Existe un equilibrio entre la atención hospitalaria y la atención ambulatoria (parámetro de "lundpararectrail"), y tanto los contactos ambulatorios como los de internación son bastante altos. La proporción de pacientes a largo plazo que permanecen en un hospital psiquiátrico durante 10 años es la más alta, lo que muestra una función del hospital psiquiátrico dirigida a la atención a largo plazo que debería proporcionarse mejor en las instalaciones residenciales. La prestación de servicios en términos de atención comunitaria se está desarrollando, como lo demuestran los altos contactos ambulatorios. Esta configuración, aunque más orientada a la comunidad que otros tipos, no muestra un patrón de .[eq]uilibrio de la atención"sino un sistema de cuidado mixto donde el desarrollo de la atención ambulatoria y primaria coexiste con la atención hospitalaria para grupos de población específicos.".* La clase C6 se define como: ***"Sistemas de salud mental bien desarrollados."***

C22 (Albania, Azerbaijan, Dominican Rep., Ukraine, El Salvador, Georgia, Guatemala, Kosovo, Moldova, Nicaragua, Paraguay, Ukraine, Uzbekistan): *"Estos países se encuentran en el grupo de ingresos medios bajos (excepto Uzbekistán, que tiene un bajo nivel de ingresos). Esta clase contiene el conjunto de todos los an-*

tiguos países soviéticos y los latinoamericanos con menos recursos. Relativamente alta disponibilidad de recursos humanos (mayor que otras clases excepto "Latvia"), pero un número bajo de hospitales psiquiátricos aunque con los gastos relativos más altos en estos hospitales a pesar de que el gasto en salud mental per cápita es bajo. La cobertura de tratamiento de la esquizofrenia es moderado. Una tendencia a los valores promedio en el parámetro "lundpararectrail" y un porcentaje bastante alto de pacientes a largo plazo que permanecen en un hospital psiquiátrico 10 años o más. Las camas aún se concentran en las principales ciudades, más que C6 y C21, aunque menos que en otros grupos. Algunos indicadores de la atención comunitaria están creciendo (existen documentos de políticas o planes, algunos de ellos también tienen legislación, pero la mayoría de ellos no tienen un informe gubernamental en cuanto a la salud mental). Esto muestra algún desarrollo del sistema de salud mental.". La definición de esta clase es: ***"Sistemas de bajos recursos basados en la institucionalidad."***

C21 (Mongolia, Hunan (China), Philippines, Vietnam): *"Estos países, en términos de recursos se encuentran en el grupo de ingresos bajos y de ingresos medios bajos. Muestran un gasto muy bajo en salud mental per cápita, mientras que el gasto relativo en hospitales psiquiátricos no es tan bajo, siendo mayor que C6 o C20. La disponibilidad de recursos humanos es inadecuada, pero casi todos los pocos recursos disponibles están en los hospitales psiquiátricos. La atención ambulatoria está poco desarrollada. La cobertura del tratamiento de la esquizofrenia es más alta que C19. "lundpararectrail" ligeramente más alto que otras clases, como C19. Menor concentración que en C19 de camas hospitalarias / psiquiátricas cerca de la ciudad más grande.La prestación de servicios de este sistema de salud es frágil: aunque hay planes o políticas disponibles, a menudo falta una legislación específica. El sistema de información no produce informes, pero todos los países tienen manuales para el tratamiento en atención primaria"*. esta clase se describe como: ***"Sistemas de salud mental enfocados en la hospitalización, sin atención ambulatoria"***

C20 (Egypt, Iraq, Iran, Morocco, West Bank, Tunisia): *"Estos países, en términos de recursos se encuentran en el grupo de ingresos medios bajos. Los sistemas de salud mental en estos países muestran aspectos contradictorios: por un lado, la atención ambulatoria se está desarrollando (este grupo tiene las cifras más altas del indicador de "lundpararectrail"); por otro lado, este grupo muestra el porcenta-*

je más alto de pacientes tratados en hospitales psiquiátricos, aunque el número de pacientes que permanecen en los hospitales psiquiátricos durante 10 años es intermedio. La concentración de camas hospitalarias alrededor de la ciudad más grande no es tan alta. El porcentaje de gastos en hospitales psiquiátricos es el más bajo entre las clases con cuidados limitados. Todos los países tienen un plan o política. Solo algunos tienen legislación sobre salud mental. Es probable que sea un grupo prometedor en términos de posibilidades de ir hacia la atención comunitaria. Según la opinión de los expertos, la existencia de asociaciones familiares y planes y políticas son indicadores de esta transición. Los expertos con conocimiento local sobre estos países informaron que la urbanización crece rápidamente, la atención comunitaria sigue siendo débil y la construcción de hospitales es la solución de emergencia para el crecimiento inminente de la población.". Con estos datos, se define a esta clase como: ***"Sistemas de salud en el borde entre cuidado comunitario y cuidado institucionalizado"***

C19 (Bangladesh, Nepal, Sri Lanka, Bhutan, India-Uttarane, Maldives, Sri Lanka, Thailand, Timor Leste): *"Estos países se encuentran en el grupo de ingresos bajos y de ingresos medios bajos. Recursos humanos bajos, cobertura de tratamiento de la esquizofrenia muy baja, el valor más pequeño del parámetro de "lundpararectrail" (excepto "Latvia"). Gastos relativamente altos en los hospitales psiquiátricos. Las camas psiquiátricas ubicadas en o cerca de la ciudad más grande son más altas que otras clases, excepto C18. Con un estado promedio de descentralización de recursos (en lo que respecta a las cifras de camas para pacientes hospitalizados cercanos a la ciudad principal), la estructura para la prestación de servicios no está bien desarrollada, pero no es tan pobre como en otros sistemas de salud mental (C18) a menudo carecen de legislación, pero algunos países tienen documentos de políticas o planes y la mayoría de ellos tienen informes del gobierno y manuales para la atención primaria. Este grupo parece representar a los países que estaban en escasez y, como no tenían una estructura hospitalaria bien establecida, evolucionan con otras fórmulas, como tratar de desarrollar la salud mental en la atención primaria. Un experto local confirma que esos países buscan alternativas, ya que no pueden costear la construcción de hospitales psiquiátricos, pasando a sistemas descentralizados no basados en la comunidad"*. Estos sistemas se definen como: ***"Sistemas de salud que funcionan desde la escazes"***.

C18 (Afghanistan, Burundi, Congo, Ethipia, Nigeria, South Africa, Uganda):

"Estos países se encuentran principalmente en el grupo de bajos ingresos, aunque el gasto en hospitales psiquiátricos es elevado. Presentan un nivel favorable para los indicadores de "lundpararectrail" (más atención ambulatoria), esto se debe a la escasez general de todos los servicios (incluida la atención hospitalaria).". Estos paises se definen como ***"Sistemas de salud mental básicos."***

3.4. Aplicación del termómetro para variables numéricas

El modelado de los termómetros para variables cuantitativas o numéricas está documentado con detalle en [Canudes Solans, 2016], sin embargo, aquí se hace un resumen de este proceso.

En la figura 3.5 se puede observar el modelo conceptual del experto sobre los datos de la OMS. El concepto que se utiliza para decidir la polaridad de la variable es el nivel de desarrollo de los sistemas de salud mental, en donde, como ya se ha mensionado, sistemas con más recursos, orientados a la inclusión social y a la atención domiciliaria de los pacientes mentales son considerados más desarrollados, mientras que sistemas con menos recursos y orientados a la hospitalización o reclusión de los pacientes mentales son considerados menos desarrollados. En este contexto, valores bajos en el parámetro *"lundpararectrail"* representan sistemas con preferencia a la hospitalización de los pacientes sobre la atención domiciliaria lo cual indica menos desarrollo por lo que se pinta de rojo. Siguiendo este concepto, valores bajos en la variable *"d2f6i71mhrec10y"* son pintados de verde puesto que representa un número pequeño de pacientes hospitalizados por 10 años o más, lo que significa que el sistema está más desarrollado y orientado a la inclusión social.

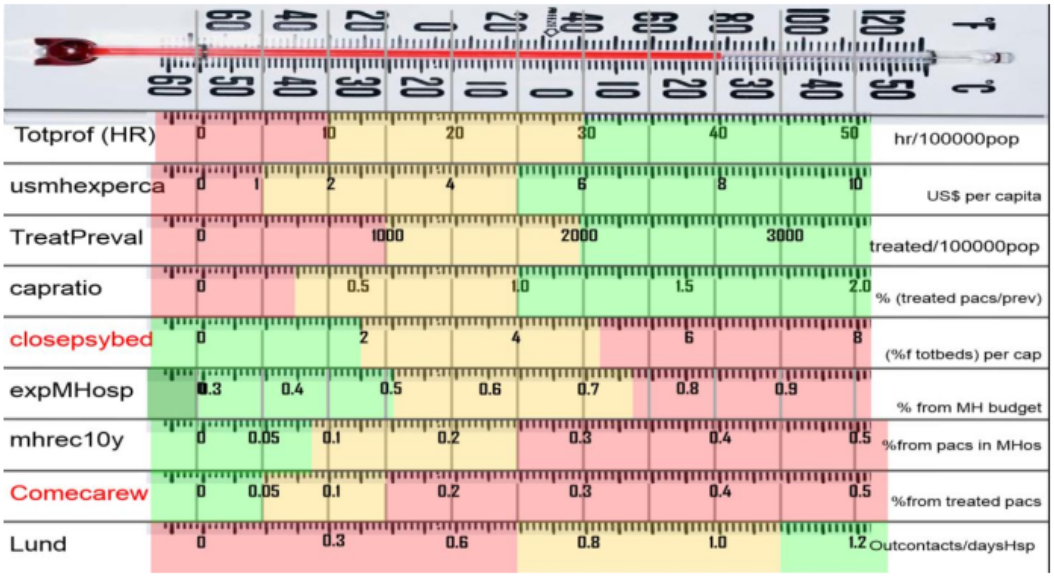

Figura 3.5: *Modelo conceptual de los termómetros para variables numéricas.* ***Fuente:*** *[Canudes Solans, 2016]*

A partir de este modelo, se ha creado el termómetro en Java-KLASS de la figura 3.6 y se ha generado el TLP que se muestra en la figura 3.7. A partir de este TLP se puede hacer una descripción de los perfiles parecida a la realizada en la sección anterior aunque solamente basada en las variables numéricas, el siguiente paso será introducir las variables categóricas al termómetro con la generalización realizada en este proyecto y verificar los resultados.

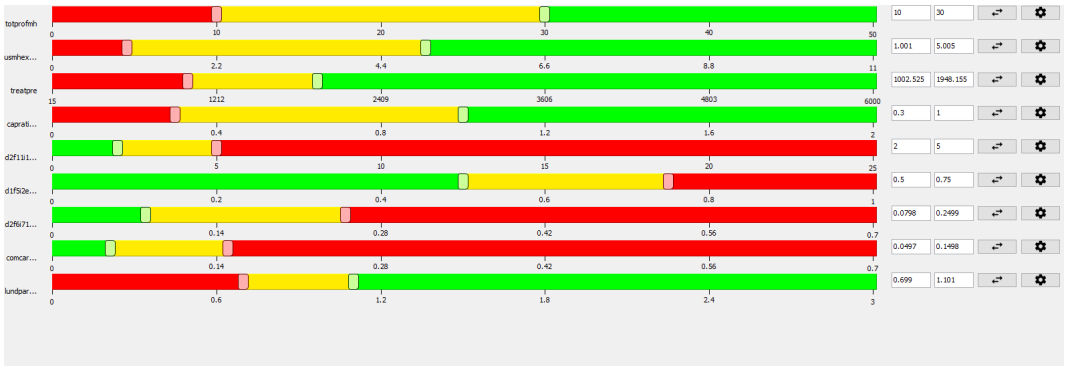

Figura 3.6: *Termómetro creado en Java-KLASS.*

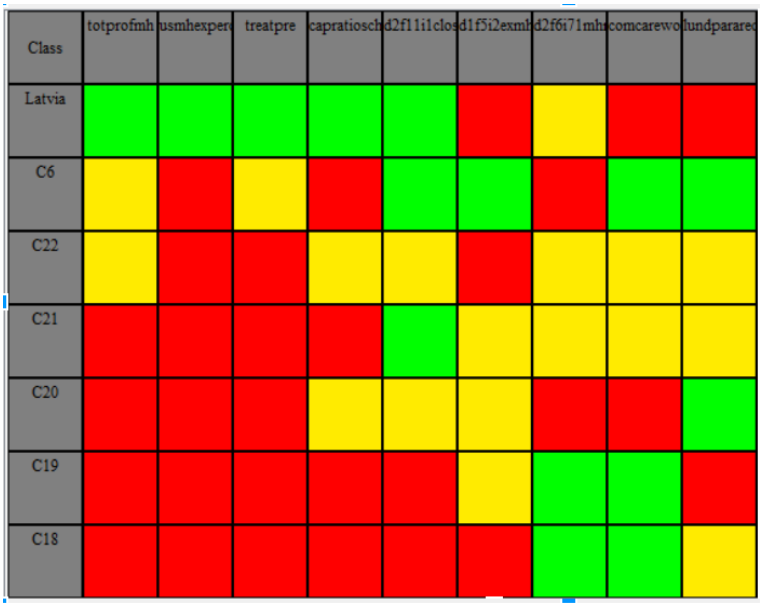

Figura 3.7: *TLP generado a partir de los termómetros para variables numéricas.*

3.5. Aplicación del termómetro para variables cualitativas

Al igual que con las variables numéricas, el concepto principal para dar un color a cada modalidad de las variables categóricas es el nivel de desarrollo de los sistemas de salud mental, de aquí, sistemas en países con mayor nivel de ingresos económicos, con presencia de legislación y planes de salud mental para los ciudadanos son considerados más desarrollados, mientras que, los países con menor ingreso económico, sin una legislación ni planes para la salud mental para los ciudadanos son considerados menos desarrollados. En la figura 3.8 se puede ver el modelo conceptual que el experto ha proporcionado para los datos de la OMS para las variables cualitativas. Aquí, la modalidad "LOW" o bajo en la variable *"Incgroup"* representa a un país con pocos ingresos económicos por lo que está pintada de rojo, por otra parte, la presencia de leyes y de planes de salud mental representan sistemas más desarrollados por lo que las modalidades "Present" de las variables *"Legisl"* y *"Polplanr"* están coloreadas de verde.

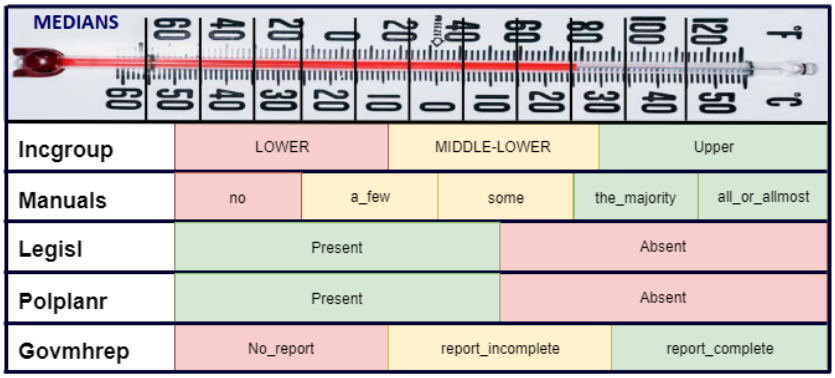

Figura 3.8: *Modelo conceptual de los termómetros para variables cualitativas.*

3.6. Creación de los termómetros en Java-KLASS

A partir de los modelos conceptuales, se introduce este conocimiento en el módulo de termómetros de Java-KLASS y se crean los termómetros para todas las variables descritas anteriormente. En la figura 3.9 se puede observar el panel con todos los termómetros creados por el usuario, los termómetros para variables numéricas permiten establecer los límites para cada color en el rango de la variable, mientras que los termómetros para variables cualitativas permite establecer un color para cada modalidad.

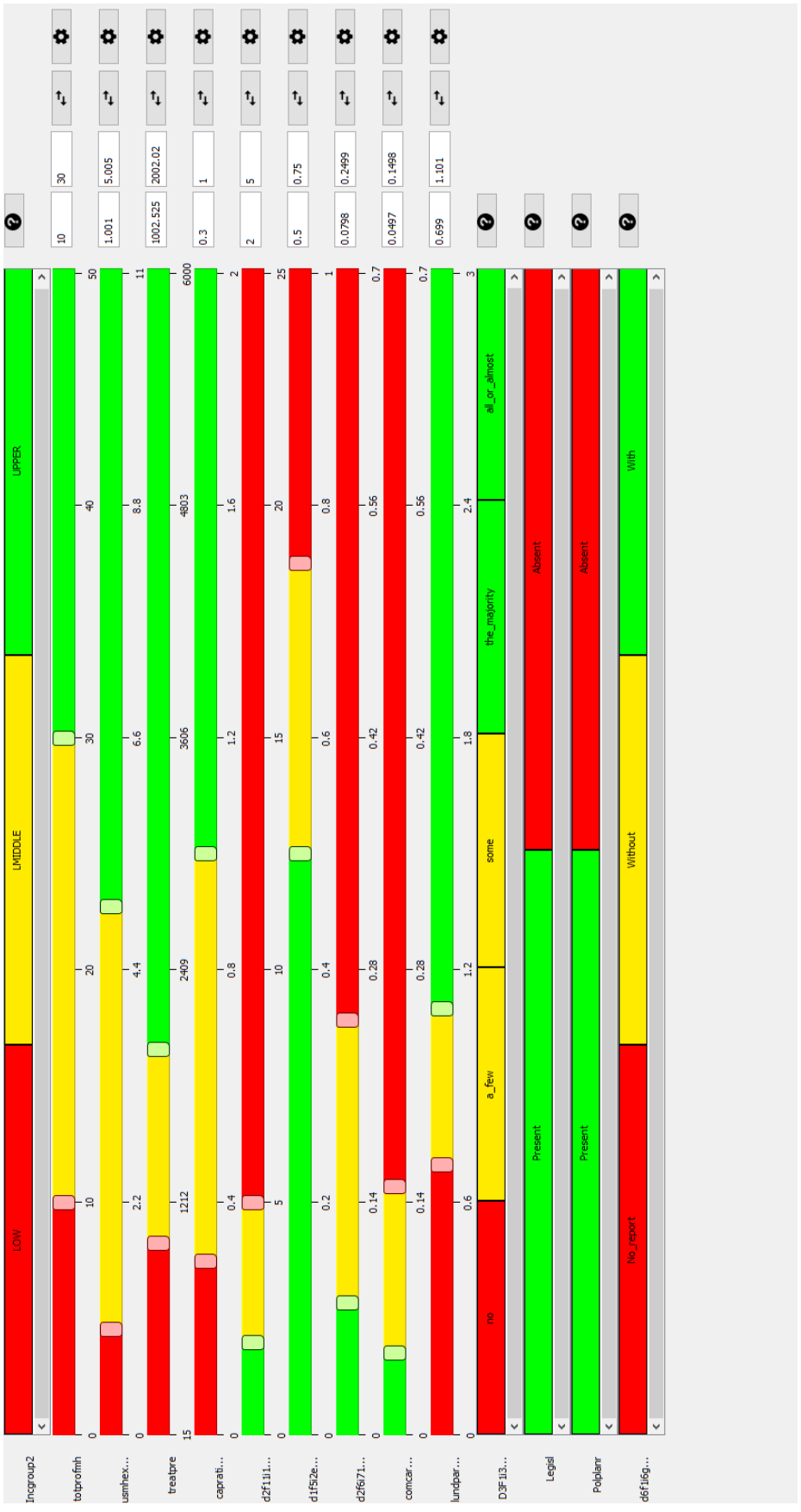

Figura 3.9: *Termómetros creados en Java-KLASS*

3.7. Construcción del TLP a partir del termómetro completo

Con la información almacenada en el termómetro anterior, se puede generar automáticamente el TLP, para esto, se utilizarán las clases descubiertas en el proceso del análisis cluster y se las asociará con las variables que se ha almacenado en el panel de termómetro. A pesar de que el proceso de generación del TLP es transparente para el usuario, aquí se mostrarán todas las fases de la generación para 4 variables de ejemplo *"Incgroup"*, *"totprofmh"*, *"usmhexperca"* y *"legisl"* siendo dos categóricas y dos numéricas. Finalmente se mostrará el cuadro semáforo completo.

1. **Fase de discretización/recodificación:** En la figura 3.10 se muestran las variables recodificadas o discretizadas según su tipo a partir del termómetro en Java-KLASS. En las 4 primeras columnas de la imagen se pueden ver los valores de las variables originales y las 4 siguientes columnas representan las variables recodificadas o discretizadas en el mismo orden.

Matriu de dades

	Incgroup2	totprofmh	usmhexperca	Legisl	VAR0	termoNum...	termoNum...	VAR1
Afghanistan	LOW	0.4	0.000081	Present	r	r	r	g
Albania	LMIDDLE	11.32	1.806866	Present	y	y	y	g
Azerbaijan	LMIDDLE	14.78	0.346054	Present	y	y	r	g
Bangladesh	LOW	0.46	0.011141	Absent	r	r	r	r
Bhutan	LMIDDLE	1.93	0.308655	Absent	y	r	r	r
Burundi	LOW	1.38	0.000007	Absent	r	r	r	r
Chile	UPPER	24.15	0.0032	Present	g	y	r	g
China_Hunan	LMIDDLE	13.5507	0.071589	Absent	y	y	r	r
Congo	LMIDDLE	1.33	0.005938	Absent	y	r	r	r
Dominican_Repu	LMIDDLE	7.59	0.126079	Present	y	r	r	g
Egypt	LMIDDLE	4.53	0.294182	Present	y	r	r	g
El_Salvador	LMIDDLE	7.96	0.040965	Absent	y	r	r	r
Eritrea	LOW	0.47	0.01	Absent	r	r	r	r
Ethiopia	LOW	1.14	0.017921	Absent	r	r	r	r
Georgia	LMIDDLE	17.76	0.828949	Present	y	y	r	g
Guatemala	LMIDDLE	2.43	0.250525	Absent	y	r	r	r
India_Uttaranc	LOW	5.1017	0.316291	Present	r	r	r	g
Iran	LMIDDLE	22.81	0.00015	Absent	y	y	r	r
Iraq	LMIDDLE	1.05	0.4102	Present	y	r	r	g
Kosovo	LMIDDLE	12.84	0.977034	Absent	y	y	r	r
Latvia	UPPER	47.23	10.172008	Present	g	g	g	g
Maldives	LMIDDLE	2.76	0.2466	Absent	y	r	r	r
Moldova	LMIDDLE	22.16	0.453685	Present	y	y	r	g
Mongolia	LOW	13.97	0.255996	Present	r	y	r	g
Morocco	LMIDDLE	4.1	0.158037	Present	y	r	r	g

Figura 3.10: *Fase de recodificación/discretización*

2. **Fase de creación de tablas cruzadas:** En la figura 3.11 se pueden ver las tablas cruzadas generadas con la herramienta de análisis bivariante de de

Java-KLASS, se han marcado los máximos de cada fila, en caso de existir empate se marca más de una celda.

Mincgroup

Clase \ VAR0	r	g	y	útils	mancants
Latvia	0	1	0	1	0
C6	0	3	0	3	0
C22	1	0	11	12	0
C21	2	0	2	4	0
C20	0	0	6	6	0
C19	4	0	4	8	0
C18	6	1	1	8	0
útils	13	5	24	42	
mancants	0	0	0		0

Mtotprofmh

Clase \ termoNumCualistotprofmh0	r	y	g	útils	mancants
Latvia	0	0	1	1	0
C6	0	3	0	3	0
C22	4	7	1	12	0
C21	2	2	0	4	0
C20	5	1	0	6	0
C19	7	1	0	8	0
C18	7	1	0	8	0
útils	25	15	2	42	
mancants	0	0	0		0

Musmhexperca

Clase \ termoNumCualisusmhexperca0	r	y	g	útils	mancants
Latvia	0	0	1	1	0
C6	1	1	1	3	0
C22	11	1	0	12	0
C21	4	0	0	4	0
C20	6	0	0	6	0
C19	8	0	0	8	0
C18	7	1	0	8	0
útils	37	3	2	42	
mancants	0	0	0		0

Mlegisl

Clase \ VAR1	r	g	y	útils	mancants
Latvia	0	1	0	1	0
C6	0	3	0	3	0
C22	5	7	0	12	0
C21	3	1	0	4	0
C20	2	4	0	6	0
C19	6	2	0	8	0
C18	4	4	0	8	0
útils	20	22	0	42	
mancants	0	0	0		0

Figura 3.11: *Tablas cruzadas*

3. **Fase de asignación del color:** En la figura se observa el TLP construido para las variables de ejemplo. En este caso no se ha ingresado un valor de gamma (γ) por lo que se tomará el valor por defecto de 0.5. Además, se puede evidenciar que se han gestionado los empates y para las filas en donde existía más de un valor igual al máximo se ha dado prioridad al amarillo.

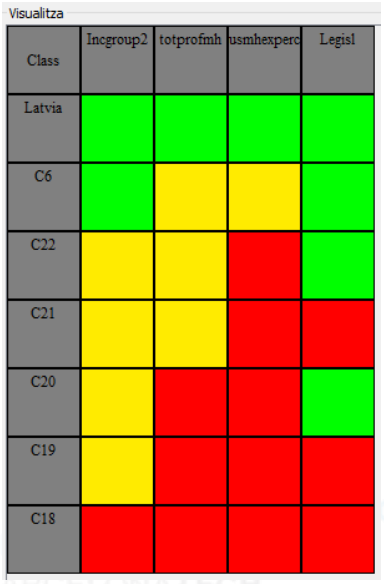

Figura 3.12: *TLP generado*

4. **Generación del TLP completo**: En la figura se puede observar el TLP basado en el termómetro, construido con todas las variables que se han descrito para la interpretación en la tabla 3.2. En este caso se ha introducido un valor de gamma (γ) de 0.75

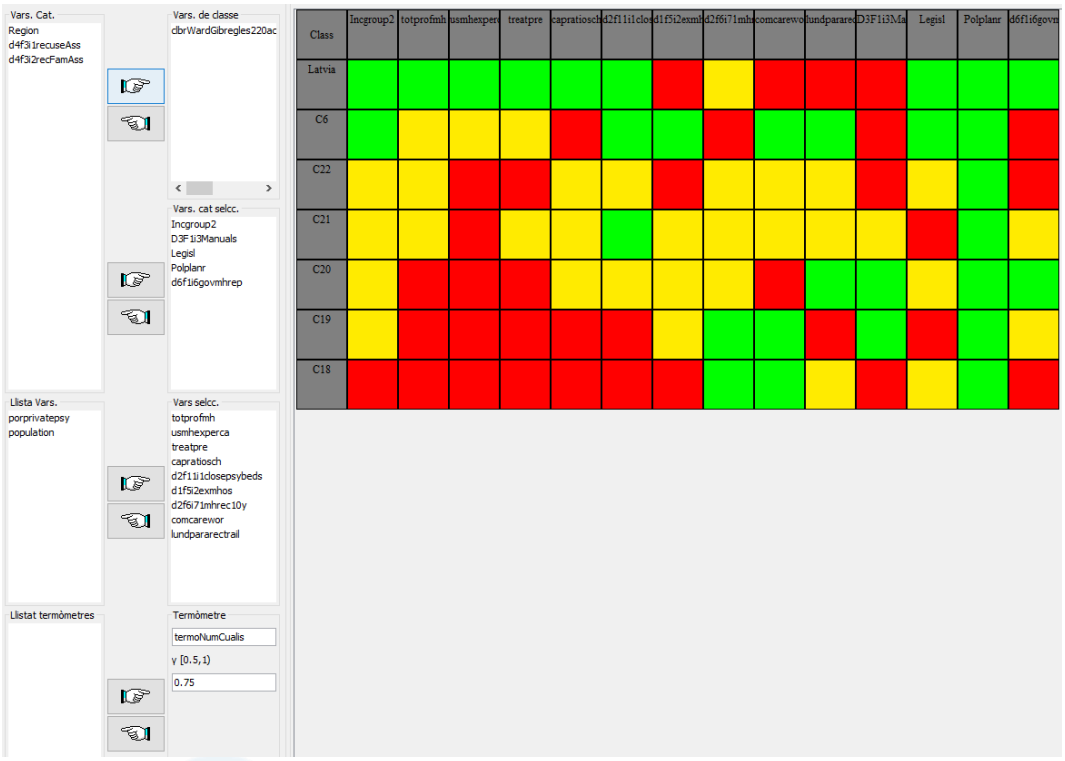

Figura 3.13: *TLP generado a partir de los termómetros en Java-KLASS con* $\gamma = 0{,}75$

En la figura 3.14 se muestra las estadísticas básicas por clases generado con la herramienta de análisis descriptivo de Java-KLASS, los valores resaltados con amarillo muestran los empates en la asignación de color y se puede comprobar que en el TLP para esas casillas se ha asignado el color amarillo. Asimismo, se ha asignado el amarillo en las celdas del TLP que representan a las variables cualitativas binarias que no superan el umbral de gamma (γ), esto se puede ver en los valores resaltados con verde del análisis descriptivo.

	CLASSE	Latvia			C6			C22			C21			C20		
	N = 42	$n_c = 1$			$n_c = 3$			$n_c = 12$			$n_c = 4$			$n_c = 6$		
VARIABLE		n_i	f_i	N*	n_i	f_i	N*	n_i	f_i	N*	n_i	f_i	N*	n_i	f_i	N*
Incgroup2	LOW	0	0	0	0	0	0	1	0.0833	0	2	0.5	0	0	0	0
	LMIDDLE	0	0	0	0	0	0	11	0.9167	0	2	0.5	0	6	1	0
	UPPER	1	1	0	3	1	0	0	0	0	0	0	0	0	0	0
Legisl	Present	1	1	0	3	1	0	7	0.5833	0	1	0.25	0	4	0.6667	0
	Absent	0	0	0	0	0	0	5	0.4167	0	3	0.75	0	2	0.3333	0
Polplanr	Present	1	1	0	3	1	0	10	0.8333	0	4	1	0	6	1	0
	Absent	0	0	0	0	0	0	2	0.1667	0	0	0	0	0	0	0

	CLASSE	C19			C18		
	N = 42	$n_c = 8$			$n_c = 8$		
VARIABLE		n_i	f_i	N*	n_i	f_i	N*
Incgroup2	LOW	4	0.5	0	6	0.75	0
	LMIDDLE	4	0.5	0	1	0.125	0
	UPPER	0	0	0	1	0.125	0
Legisl	Present	2	0.25	0	4	0.5	0
	Absent	6	0.75	0	4	0.5	0
Polplanr	Present	6	0.75	0	6	0.75	0
	Absent	2	0.25	0	2	0.25	0

Figura 3.14: *Análisis descriptivo por clases*

3.8. Comparación con el TLP original

En esta sección se hace una comparativa del TLP original construido manualmente con ayuda de los expertos que se muestra en la sección 3.1 con los TLPs generados automáticamente con Java-KLASS. Para ello se han creado los siguientes TLP basados en termómetros:

- **I:** TLP basado en termómetros para variables numéricas sin gestión de empates.
- **II:** TLP basado en termómetros para variables numéricas con gestión de empates.
- **III:** TLP basado en termómetros completo con gestión de empates con gamma=0.75.
- **IV:** TLP basado en termómetros completo con gestión de empates con gamma=0.50.
- **V:** TLP basado en termómetros completo con gestión de empates con gamma=0.85.

En la figura 3.15 se muestra las configuraciones usadas para cada TLP, se han marcado con color azul las celdas que presentan diferencias con el TLP original. Luego se muestra una tabla comparativa 3.3 en donde se puede observar que el caso II que incluye el método de gestión de empates ha sido beneficioso par los termómetros de variables numéricas, aumentando el porcentaje de celdas y variables iguales respecto al original. Al incluir las variables cualitativas aumenta el número de celdas diferentes, aunque 3 de las variables cualitativas aparentan ser robustas, para las otras 2 variables cualitativas binarias, la que más se acerca al original es la que tiene un valor de gamma (γ) de 0.85, pero se está estudiando la posibilidad de calcular el gamma (γ) óptimo para cada variable cualitativa independiente, lo que da lugar al caso `VII`.

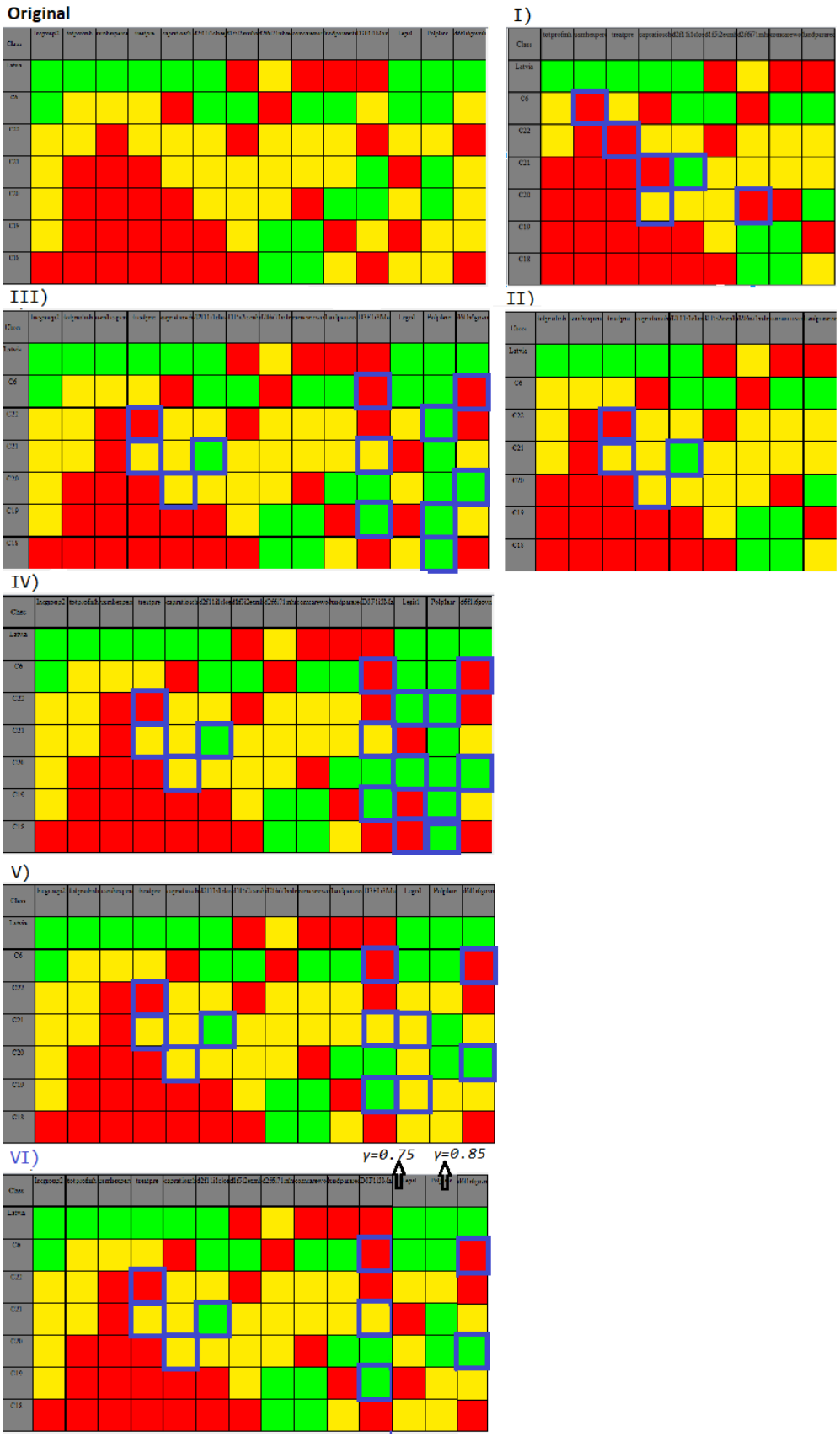

Figura 3.15: *Comparación de TLPs generados con diferentes configuraciones*

Comparación	TLP Comparado					
	I	II	III	VI	V	VI
Celdas de diferente color	6	4	12	15	11	9
Variables diferentes	5	3	6	7	6	5
Celdas Comparadas	56	56	98	98	98	98
Variables Comparadas	9	9	14	14	14	14
% de celdas iguales	89	92	88	84	89	92
% de variabless iguales	44	66	57	50	57	64

Tabla 3.3: *Tabla de comparación de TLPs respecto al original basado en expertos*

No obstante, con el semáforo anotado se puede obtener una información adicional para valorar si estas diferencias son muy relevantes o no en función de la pureza de la clase.

3.8.1. Comparación entre TLPs anotados

Como se puede observar el TLP completo que mejor se ajusta al original es el que tiene un gamma (γ) de 0.85, a continuación se anota este TLP de la forma que se explica en la sección 1.4.2 y se vuelve a comparar con el TLP original también anotado. En la figura 3.16 se muestran los dos aTLPs en los que se han señalado con azul las celdas diferentes. Puede observarse que en todos los casos la celdas distintas están obscurecidas, esto quiere decir que son celdas con ruido o, expresado formalmente, presentan unos CVs elevados. Sin embargo, para algunas celdas como por ejemplo en la variable "legisl" para las clases C21 y C19, el aTLP generado a partir del termómetro presenta colores más puros en las celdas diferentes, esto también se evidencia con claridad en la variable "treatpre" (columna 4) para la clase C21 (fila 4), esto quiere decir que el TLP basado en termómetros presenta colores que se acercan más a la tendencia central localmente mayoritaria de las variables asociadas en cada clase.

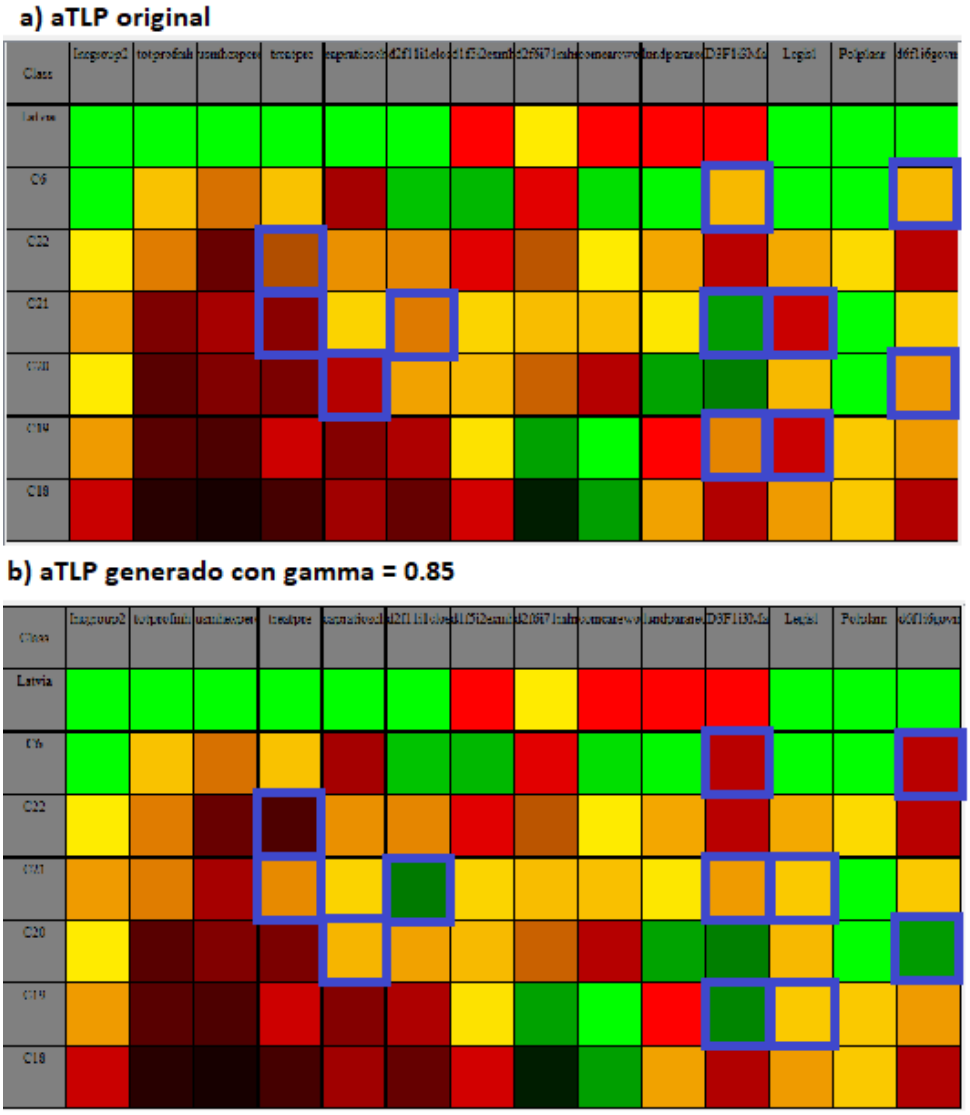

Figura 3.16: *Comparación del aTLP original con el generado con* $\gamma = 0{,}85$

3.8.2. Verificación de los perfiles a partir del TLP generado

A partir de la descripción de los perfiles de los sistemas de salud mental descritos para cada clase en la sección 3.1, a continuación se analiza cada uno de ellos y se discute que no deberían estar en la descripción o que conceptos deberían incluirse de acuerdo al TLP generado automáticamente en este proyecto.

Clase Latvia: La clase "Latvia" no tiene cambios ya que todas las celdas del TLP son iguales.

C6: En esta clase se ha dicho que son sistemas de cuidados mixtos donde la atención ambulatoria y primaria coexiste con la atención hospitalaria, sin embargo la variable "D3f1i3Manuals" muestra valores bajos(se ha coloreado de rojo) por lo que se puede decir que la atención primaria está menos desarrollada que la atención ambulatoria y hospitalaria. De igual forma se puede agregar estos sistemas no presentan reportes al departamento de salud mental de los gobiernos.

C22: En esta clase el parámetro "treatpre" se ha pintado de rojo por lo que se puede agregar que la cobertura de atención a los pacientes mentales es deficiente. Las demás variables no presentan diferencias.

C21: En el perfil de esta clase se ha dicho anteriormente que el número de recursos humanos es inadecuado, sin embargo, este parámetro se ha pintado de amarillo lo que significa que está a la altura de las clases C6 y C22. De igual forma, la cobertura de atención a los pacientes mentales es superior al de la clase C22; anteriormente se había dicho lo contrario.

C20: En esta clase se puede agregar que los sistemas de salud mental no tienen una adecuada cobertura al tratamiento de la esquizofrenia.

C19, C18: Estas clases no presentan cambios en cuanto a su coloración, por lo tanto la interpretación es la misma.

3.9. Discusión

Determinar el valor de gamma (γ) adolece de los mismos problemas que corresponden a todos los algoritmos que tienen parámetros de entrada, que es, encontrar un buen criterio para determinar con que valor de gamma (γ) para construir los semáforos. Gamma (γ) es fácil de interpretar porque tiene que ver con el ruido que aceptamos en una clase para asignarla un color determinado, esto intuitivamente es fácil de entender y fácil determinar gamma de acuerdo al coste de la asignación equivocada. En un problema real cualquiera un experto sería capaz de decir hasta que tolerancia podría aceptar de variación dentro de la clase respecto al color asignado para él poder asumir los costes del tratamiento equivocado de los elementos que se asignan un color diferente. El hecho de que las gammas más adecuadas pue-

dan ser distintas en una variable o en otra nos ponen en una situación de tener que determinar estas gammas con un criterio que ya no sea general para todo el caso de estudio, entonces se plantea una línea de investigación futura que es ver con que criterio se podría determinar gamma de forma automática utilizando solamente información interna de la muestra, ya que, la variable o los resultados con los que se están comparando ahora, que han servido para hacer benchmarking, en una aplicación real no se tendrán, serán exactamente el resultado que se intenta obtener, por lo que no se puede asignar gamma como si se estuviera en un entorno supervisado por lo tanto sería interesante ver como la información interna de las variaciones internas de cada una de las clases crean criterios para asignar una gamma local a cada una de las variables cualitativas.

Capítulo 4

Conclusiones

4.1. Conclusiones

A lo largo de este proyecto, se ha analizado el proceso de KDD con un énfasis principal en la fase de interpretación de los resultados y generación del conocimiento. Para esto, en primer lugar se ha realizado una revisión del estado del arte y se han presentado tres herramientas que ayudan a la interpretación de las clases descubiertas al ejecutar un proceso de minería de datos como son el CPG, el TLP y el aTLP. De igual forma se ha introducido el termómetro como una herramienta que permite introducir el conocimiento a priori de los expertos para transferir la polaridad semántica de las variables al TLP. Luego se ha hecho una breve explicación del proceso de KDD y una introducción a la herramienta java-Klass como un software que brinda un conjunto de herramientas para ayudar a los expertos en el proceso de minería de datos.

A continuación, este documento se ha enfocado principalmente en describir el modelo de termómetro propuesto originalmente en [Canudes Solans, 2016] para variables numéricas y el proceso de generación automática del TLP a partir de estos; siguiendo con esta línea, se ha propuesto un moldeamiento visual y estructural para extender los termómetros a variables cualitativas así como las modificaciones al proceso de generación automática del TLP para que sea posible generarlo con el termómetro extendido. Al final se ha presentado un caso practico en donde se analizan los datos de la OMS respecto a los sistemas de salud mental en países en vías de desarrollo. Finalizando este proyecto se ha comprobado que los objetivos específicos planteados al principio se han cumplido completamente como se describe a continuación:

1. *Extender los termómetros a variables cuantitativas:* Se ha presentado el modelo visual y estructural para los termómetros para variables cualitativas así como las modificaciones realizadas en el diagrama de clases de java para que se generalicen los termómetros y se ha comprobado que el sistema no pierda la funcionalidad original y que a la vez sea capás de funcionar con las nuevas características de los termómetros para variables cualitativas.

2. *Crear un método de recodificación de variables cualitativas basado en termómetros que asigne un código a las modalidades de la variable según los colores del termómetro:* El algoritmo de recodificación propuesto funciona de manera similar a una asignación directa en el que se crea un código de color de acuerdo al color asignado manualmente a cada modalidad de la variable cualitativa en el termómetro.

3. *Flexibilizar la implementación del TLP basado en termómetros actual para que: a) admita termómetros que incluyan variables cualitativas, b) permita trabajar con termómetros que no incluyan todas las variables del TLP o puedan contener variables adicionales:* Las modificaciones realizadas al proceso de generación del TLP original periten cumplir con estos dos requerimientos, además de esto, se ha incluido la gestión de empates como un modelo conservador que asigna el amarillo en el caso de existir empates en la asignación de color evitando así la aleatoriedad del modelo anterior en estos casos. Asimismo, se ha incluido el parámetro gamma (γ) a las variables cualitativas binarias para que sea posible generar un semáforo con los tres colores a partir de una variable con únicamente dos modalidades.

4. *Comparar el TLP generado automáticamente a partir de los termómetros generalizados con el TLP que se crearía de forma manual con ayuda del experto a partir de las distribuciones condicionales de las variables respecto a las clases:* En el caso de estudio se ha realizado la comparación del TLP creado manualmente con ayuda de los expertos a partir del CPG y un análisis estadístico básico de las variables asociadas, con varios TLPs con distintas configuraciones que se han generado automáticamente a partir del termómetro para variables cualitativas y cuantitativas. Aquí se ha observado que las modificaciones realizadas para la gestión de los empates aproximan de mejor manera el TLP generado al original, y que en general la construcción de el TLP basado en termómetros brinda una aproximación bastante buena al TLP original. De igual forma se evidencia como el parámetro gamma (γ) sirve

para ajustar mejor las variables cualitativas binarias. Al anotar los TLPs se evidencia que el TLP que se ha construido a partir del termómetro presenta colores más brillantes en las celdas que presentan diferencias con el TLP original, esto quiere decir que el TLP generado automáticamente a partir del termómetro ajusta mejor a las tendencias centrales de las variables en cada clase.

Otra conclusión importante a la que se ha llegado con del caso de estudio, es que, a partir de que se han extendido los semáforos y los termómetros ahora se puede trabajar transmitiendo esta semántica original que los expertos le dan a las variables directamente en el semáforo. Que el semáforo sea un elemento de interpretación de clases que directamente está en consonancia con la conceptualización que tiene el experto de las variables que está manejando, esto acerca a la herramienta a los códigos de interpretación del experto y por tanto facilita que la compresión de las clases sea más directa.

A parte de que el proyecto ha servido para esto, otra cosa importante es que se ha visto, de alguna manera, como las diferentes configuraciones que se han estudiado realmente reproducen bastante bien lo que habían hecho los expertos manualmente al principio. El semáforo original estudiado ha sido construido exclusivamente observando las distribuciones condicionadas que aparecen en el CPG, y por tanto, con un criterio bastante más intuitivo; entonces, constatar que con elementos de tipo más computacional o más científicos y criterios objetivables se llega a reproducir bastante bien algo que está expresando la intuición del experto es una contribución de esta tesis que tiene repercusiones en los procesos de interpretación automática de clases.

4.2. Futuras líneas de investigación

Los siguientes trabajos o lineas de investigación pueden ser abordados en el futuro en el marco de este proyecto:

1. Al poder realizar una interpretación visual completa de las clases con el TLP generado partir de los termómetros extendidos, se puede hacer un análisis comparativo con otras herramientas de conceptualización de clases que pre-

senta java-KLASS como la inducción de reglas basada en boxplots o la conceptualización jerárquica (CCEC).

2. Como se menciona en la sección 3.9, al realizar la comparación con del TLP original con el generado en el análisis del caso práctico se ha evidenciado que se puede necesitar un ajuste diferente en el parámetro gamma (γ) para cada variable, por lo tanto, en un futuro se plantea una línea de investigación futura que es ver con que criterio se podría determinar gamma de forma automática utilizando solamente información interna de las variaciones internas de cada una de las clases para crear criterios con el fin de asignar una gamma local a cada una de las variables cualitativas.

ACRÓNIMOS

aTLP *annotated Traffic Ligth Panel*. 5, 6, 63, 67, 71

BV *Base de Conocimiento*. 17, 71

CPG *Class Panel Graph*. 4, 5, 15, 16, 23, 30, 46, 67–69, 71

CV *Coeficientes de Variación*. 5, 6, 63, 71

FIB *Facultat d'Informàtica de Barcelona*. 15, 71

KB *Knowledge Base*. 4, 71

KDD *Knowledge Discovery in Databases*. 4, 9, 10, 14, 67, 71

LAMIC *Low and middle income countries*. 42, 71

OMS *Organización Mundial de la Salud*. 8, 42, 43, 51, 53, 67, 71

TLP *Traffic Ligth Panel*. 2–7, 15, 23, 25, 29, 30, 35, 36, 38, 39, 42, 46, 52, 56, 58–61, 63, 64, 67, 68, 71

UIB *Universitat Illes Balears*. 16, 71

UPC *Universitat Politècnica de Catalunya*. 16, 71

Bibliografía

[Bayona, 2000] Bayona, S. (2000). Descriptiva de dades y de classes. *PFC Facultad d'Informática, UPC*.

[Canudes Solans, 2016] Canudes Solans, D. (2016). Apropant el datamining a l'expert a través del pre i postprovessament de resultats: l'ús del termòmetre en la construcció automàtica dels quadres semàfors de klass. Master's thesis, Universitat Politècnica de Catalunya.

[Castillejo, 1996] Castillejo, X. (1996). Un entorn de treball per a klass. *PFC Facultat d'Informatica, UPC*, page 32.

[Fayyad et al., 1996] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34.

[Gibert, 1991] Gibert, K. (1991). Klass. estudi d'un sistema d'ajuda al tractament estadıstic de grans bases de dades. Master's thesis, Master's thesis, UPC.

[Gibert, 1995] Gibert, K. (1995). *L'us de la informació simbòlica en l'automatització del tractament estadístic de dominis poc estructurats*. PhD thesis.

[Gibert, 2014] Gibert, K. (2014). Mixed intelligent-multivariate missing imputation. *International Journal of Computer Mathematics*, 91(1):85–96.

[Gibert and Conti, 2015] Gibert, K. and Conti, D. (2015). atlp: A color-based model of uncertainty to evaluate the risk of decisions based on prototypes. *AI Communications*, 28(1):113–126.

[Gibert and Conti, 2016] Gibert, K. and Conti, D. (2016). On the understanding of profiles by means of post-processing techniques: an application to financial assets. *International Journal of Computer Mathematics*, 93(5):807–820.

[Gibert et al., 2012a] Gibert, K., Conti, D., and Sànchez-Marrè, M. (2012a). Decreasing uncertainty when interpreting profiles through the traffic lights panel. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 137–148. Springer.

[Gibert et al., 2012b] Gibert, K., Conti, D., and Vrecko, D. (2012b). Assisting the end-user in the interpretation of profiles for decision support. an application to wastewater treatment plants. *Environmental Engineering and Management Journal*, 11(5):931–944.

[Gibert and Cortés, 1997] Gibert, K. and Cortés, U. (1997). Weighting quantitative and qualitative variables in clustering methods. *Mathware & soft computing. 1997 Vol. 4 Núm. 3.*

[Gibert et al., 2010a] Gibert, K., García-Alonso, C., and Salvador-Carulla, L. (2010a). Integrating clinicians, knowledge and data: expert-based cooperative analysis in healthcare decision support. *Health research policy and systems*, 8(1):28.

[Gibert et al., 2008a] Gibert, K., García-Rudolph, A., and Rodríguez-Silva, G. (2008a). The role of kdd support-interpretation tools in the conceptualization of medical profiles: An application to neurorehabilitation. *Acta Informatica Medica*, 16(4):178.

[Gibert et al., 2018] Gibert, K., Horsburgh, J., Athanasiadis, I., and Holmes, G. (2018). Environmental data science. *Environmental Modelling and Software.*

[Gibert et al., 2008b] Gibert, K., Izquierdo, J., Holmes, G., Athanasiadis, I., Comas, J., and Sànchez-Marrè, M. (2008b). On the role of pre and post-processing in environmental data mining.

[Gibert et al., 2010b] Gibert, K., Martín, J., and . Salvador-Carulla, L. (2010b). Who-aims: General clustering. Technical part.

[Gibert et al., 2005] Gibert, K., Nonell, R., Velarde, J., and Colillas, M. (2005). Knowledge discovery with clustering: Impact of metrics and reporting phase by using klass. *Neural Network World*, 15(4):319.

[Gibert et al., 2013] Gibert, K., Rodríguez-Silva, G., and Annicchiarico, R. (2013). Post-processing: Bridging the gap between modelling and effective decision-

support. the profile assessment grid in human behaviour. *Mathematical and Computer Modelling*, 57(7-8):1633–1639.

[Gibert and Salvador, 2000] Gibert, K. and Salvador, A. (2000). Aproximación difusa a la identificación de situaciones características en el tratamiento de aguas residuales. In *X Congreso Espanol sobre tecnologıas y lógica fuzzy*, pages 497–502.

[Gibert et al., 2010c] Gibert, K., Sànchez-Marrè, M., and Codina, V. (2010c). Choosing the right data mining technique: classification of methods and intelligent recommenders. In *Proc. of the iEMSs'10, 5th biennial meeting: (III DMTES Workshop), S23.03.1-S23.03.9.*

[Gibert et al., 2016] Gibert, K., Sànchez-Marrè, M., and Izquierdo, J. (2016). A survey on pre-processing techniques: Relevant issues in the context of environmental data mining. *AI Communications*, 29(6):627–663.

[Goebel and Gruenwald, 1999] Goebel, M. and Gruenwald, L. (1999). A survey of data mining and knowledge discovery software tools. *ACM SIGKDD explorations newsletter*, 1(1):20–33.

[Hand, 2007] Hand, D. J. (2007). Principles of data mining. *Drug safety*, 30(7):621–622.

[Leiva and S, 2018] Leiva, M. and S, T. (2018). Knowledge discovery in databases. urlhttp://mnrva.io/kdd-platform.html. Accedido 15-05-2018.

[Márquez and Martın, 1997] Márquez, J. and Martın, J. (1997). La clasificación automática en las ciencias de la salud. *PFC. Facultat de Matematiques i Estadıstica, UPC.*

[Mollá Santiago, 2014] Mollá Santiago, S. (2014). Generalització de mètodes de density-based clustering a dades mixtes.

[Murtagh and Legendre, 2014] Murtagh, F. and Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *Journal of classification*, 31(3):274–295.

[Prather et al., 1997] Prather, J. C., Lobach, D. F., Goodwin, L. K., Hales, J. W., Hage, M. L., and Hammond, W. E. (1997). Medical data mining: knowledge discovery in a clinical data warehouse. In *Proceedings of the AMIA annual fall symposium*, page 101. American Medical Informatics Association.

[Rodas-Osollo, 2004] Rodas-Osollo, J. (2004). Knowledge discovery in repeated and very short serial measures with a blocking factor. *AI Communications*, 17(3):175–178.

[Tubau, 1999] Tubau, X. (1999). Sobre el comportament de les metriques mixtes en algorismes de clustering. *PFC.[cited at p. 32].*

[Vázquez and Gibert, 2002] Vázquez, F. and Gibert, K. (2002). Robustness of class prediction depending on references partition in-ill-structured domains. 8th. In *Iberoamerican Conference on Artificial Intelligence. Sevilla, España.*

[Zhu, 2007] Zhu, X. (2007). *Knowledge Discovery and Data Mining: Challenges and Realities: Challenges and Realities.* Igi Global.